Table of Contents:

Preface: Are Large Regional or Special Topic Servers Necessary?
Maurer, Hermann ................................................................. 1

Austria-Forum and Beyond
Maurer, Hermann ........................................................................ 3

Web Books: The Fusion of Paper and Pixels
Pirker, Johanna; Wurzinger, Gerhard; Müller, Heimo ................................................................. 9

Verifying Information on the Internet
Korica-Pehserl, Petra ..................................................................... 15

Geographic Data Verification - Automatic Verification of Physical Geographic Data Using Maps
Mehmood, Rizwan .......................................................................... 20

How Open Content Servers Can Be Made Beneficial for Learning and Education
Ebner, Martin; Scerbakov, Nicolai; Schön Sandra ................................................................. 26

Digital Archives as Part of Digital Humanities Research Infrastructure
Stigler, Hubert Johannes ............................................................... 31

Thematic Digital Libraries vs. Wikipedia's "One Size Fits All" - Lessons Learned
Barthel, Simon; Balke, Wolf-Tilo ....................................................... 37

Learning Management Systems - A Need for Specialized Systems
Sharif, Nauman; Afzal, Tanvir; Helic, Denis ......................................................... 44

An Educational Framework for Content Sharing
Baumgartner, Peter; Zobel, Annett ....................................................... 48

Heritage Portals and Heritage Mining: Synergizing Data and Image Mining
Under Uncertainty Constraints
Milutinovic, Veljko; Salom, Jakob; Jelisavcvic, Vladimir; Filipovic, Viktorija; Mihajlovic, Aleksandar;
Ognjanovic, Zoran; Markovic, Zoran; Kos, Anton; Tomazic, Saso; Maurer, Hermann ................. 53
The IPSI BgD Internet Research Society

The Internet Research Society is an association of people with professional interest in the field of the Internet. All members will receive these TRANSACTIONS upon payment of the annual Society membership fee of €500 (air mail printed matters delivery).

*Member copies of Transactions are for personal use only*

**IPSI BGD TRANSACTIONS ON ADVANCED RESEARCH**

www.internetjournals.net

<table>
<thead>
<tr>
<th>STAFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Veljko Milutinovic, Co-Editor-in-Chief</td>
</tr>
<tr>
<td>Department of Computer Engineering, ETF, University of Belgrade, POB 35-54, Belgrade, Serbia, Tel: (381) 64-1389281</td>
</tr>
<tr>
<td><a href="mailto:vm@eft.rs">vm@eft.rs</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EDITORIAL BOARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipkovski, Aleksandar</td>
</tr>
<tr>
<td>The Faculty of Mathematics, Belgrade, Serbia</td>
</tr>
<tr>
<td>Blaisten-Barojas, Estela</td>
</tr>
<tr>
<td>George Mason University, Fairfax, Virginia, USA</td>
</tr>
<tr>
<td>Crisp, Bob</td>
</tr>
<tr>
<td>University of Arkansas, Fayetteville, Arkansas, USA</td>
</tr>
<tr>
<td>Domenici, Andrea</td>
</tr>
<tr>
<td>University of Pisa, Pisa, Italy</td>
</tr>
<tr>
<td>Flynn, Michael</td>
</tr>
<tr>
<td>Stanford University, Palo Alto, California, USA</td>
</tr>
<tr>
<td>Fujii, Hironori</td>
</tr>
<tr>
<td>Fujii Labs, M.I.T., Tokyo, Japan</td>
</tr>
<tr>
<td>Ganascia, Jean-Luc</td>
</tr>
<tr>
<td>Paris University, Paris, France</td>
</tr>
</tbody>
</table>
Preface: Are Large Regional or Special Topic Servers Necessary?

Maurer, Hermann

Abstract - The main question addressed in this special issue consisting of 10 papers is this. Can powerful search engines, in combination with large information servers like Wikipedia and some very specialized smaller servers satisfy the need to provide information for all instances when special regions, interests or topics are to be served. As it turns out, the answer is a clear no.


1. SUMMARY

The question treated in this special issue can be formulated like this. If we have powerful search engines will a server concept like Wikipedia work for all kinds of special large information servers? To put it more sloppy: Can we have a server model that fits all wishes, or is it like with T-shirts: if we have a T-shirt that fits all it really does not fit anyone well. The answer is surprisingly clear: a server concept that fits all does not exist, yet we need different information servers for a variety of purposes. It is interesting that different applications offer completely different reasons why this is so. In what follows we briefly explain the arguments leading to above conclusion paper by paper.

2. DISCUSSION OF PAPERS

The first paper: “Austria-Forum and beyond” describes a large working server that is supposed to provide reliable and quotable information of interest to and for Austrians. Although a large German language Wikipedia exists that also covers many aspects of Austria, even if the information were tailored to cater for Austrian needs this would not suffice. The main reasons are (a) the anonymity of most Wikipedia contributions is sometimes hiding sloppiness and inaccuracies making contributions not sufficient reliable and quotable, and (b) all the wisdom available in printed books is not directly included into any Wikipedia. Thus, the Austria-Forum is unique in that it weaves together ordinary Web pages with digitized books. This idea that has led to the Serbia-Forum has made it possible to be also used as the basis for presenting Academia Europaea.

The second paper: “Web Books: The Fusion of Paper and Pixels” emphasizes that new interactive functions have to be included into a library of Web Books and points out that such digital libraries should not be seen as isolated collections of information, but should be integrated with other information sources as is attempted in Austria-Forum.

The third paper: “Verifying Information on the Internet” addresses the issue of reliability of material found in the Web. It focuses on the concept of verifying existing material by experts, a process that is explained with samples from a large working application.

The fourth paper: “Geographic Data Verification” is explaining how in some areas like in geography, the reliability of textual information can be automatically or semi-automatically verified using cartographic material. It must allow special queries”. i.e. it must prove operators to work with different cartographic features. E.g. it should allow showing result of the intersection of e.g. the curve representing a river with the map of countries, thus allowing verifying if the list of countries the river flows through is indeed correct.

Paper five: “How Open Content Servers Can Be Made Beneficial for Learning and Education” looks at servers built with the explicit aim to support Learning and Education. For educators as well as learners searching for open content on several platforms can be very exhausting. In this publication a first prototype for mobile devices is presented that allow users to find open educational resources in minutes.

Paper six: “Learning Management Systems- A need for Specialized Systems” stresses the point that also in educational environments systems have to behave differently for different topics and different groups of learners, thus very much supporting the notion that there is no e-Learning system that would fit all needs.

Preface received June 2, 2014. Hermann Maurer is Professor of Informatics at Graz University of Technology in Austria and member of the Board of Trustees of Academia Europaea, see www.ae-info.org for further details and email see http://www.ae-info.org/ae/USER/Maurer_Hermann
The seventh paper: “An Educational Framework for Content Sharing” proposes a concrete and working model for using OER (Open Educational Resources) as also discussed in paper 5. It is argued that the lackluster success of e-Learning is due to the fact that material tailored to a specific group of learners is hard to find or rests on modifying existing material, requiring Open Content instead of building up repositories organized just by content criteria. The paper proposes a community platform (edu-sharing.net) supported by a search engine triggered by criteria taken from an educational taxonomy of teaching methods.

Paper eight: “Digital archives as part of Digital Humanities research infrastructure - Towards a standardized model of archiving and dissemination” argues convincingly that all large information servers need a large amount of metadata to make them truly valuable. XML-based data formats appear well-suited to flexible, metadata-enriched forms of storage of textual data: the primary content of documents is augmented with additional descriptive elements, based on one of the many modelling standards. This particular paper presents and discusses an object-oriented approach to a digital archive, which was implemented in a FEDORA-based (Flexible Extensible Digital Object Repository Architecture) digital repository environment.

The ninth paper: “Thematic Digital Libraries vs. Wikipedia's 'One Size Fits All' - Lessons Learned” answers the question whether one type of server can handle very different topics in a diplomatic way: It argues that although one size does not fit all, there is a lot to learn from general purpose portals. However, it also shows that in many instances special approaches are needed. The paper explores the challenges and chances of specialized thematic digital libraries by reviewing typical use cases from disciplines like chemistry or mathematics.

The tenth and the final paper: “Heritage Portals and Heritage Mining: Synergizing Data and Image Mining Under Uncertainty Constraints” sheds light on the essential issues of Serbia-Forum, introduces the concept of heritage mining, and views digital preservation of national heritage as the enabler technology for heritage mining.

ACKNOWLEDGMENT

I want to thank Professor Veljko Milutinovic for suggesting that I try to edit such special issue. I would have not been able to do so without the patient support of Mr. Jakob Salom. I do hope readers will learn from reading the papers in this issue as much as I have learned from them and from my correspondence with authors and reviewers.
Austria-Forum and Beyond

Maurer, Hermann

Abstract - Austria-Forum is a substantial Web site increasingly used in addition to other online encyclopedias. In this paper we describe exactly what it is, why it is being developed, how we see its future, and why we believe it is an important model as an attempt to beat the tyranny of search engines and information servers whose contents are of unknown quality.

Index Terms: Austria-Forum, Encyclopedias, Knowledge database, Semantic Web, Wikipedia, WWW.

1. INTRODUCTION

THE Austria-Forum [1] started with a project now 20 years ago whose aim was to present an online encyclopedia of Austria, including a mix of media. The project was quite successful: the server is actually still in operation but is now more a historic relic than anything else [2]. After 1997 a number of academic experiments with both contents and technology were carried out, but no new product was developed. In 2009 a decision was necessary: to either shut the server down, or to do something much bigger. The decision was to go for the second alternative.

The scope of the new project has further expanded since then. Rather than describing the process we concentrate in the next Section 2 on what the Austria-Forum is right now and how it is run and managed. We explain what makes it interesting and different from other attempts in Section 3. We discuss some special features including problems and weaknesses in Section 4. In Section 5 we describe longer features: the realization of some will depend on whether enough financing can be found. Section 6 is a short summary, followed by a few references.

A word of caution seems to be in order: Austria-Forum is work in progress. Thus, not all features of the current systems described are fully implemented for some parts of the information.

AUSTRIA-FORUM, STATUS 2014

The Austria-Forum is a large collection of general information that is of interest for the general Austrian public subject to a number of restrictions:

(i) Material only understandable to specialists is usually not included;
(ii) Information with little interest after a short time is not found (i.e. no reports on current politics, weather, sports, or similar): this is left to other media;
(iii) Contributions should have a known source (in case of material compiled by persons typically their names and CVs, or otherwise books or archives that are clearly mentioned);
(iv) Information is “frozen” in many cases, i.e. cannot be changed any more, but comments can be added at any time. Note: If a contribution gets rather outdated or some other points of view are to be incorporated, this can be done by including a new entry properly linked to others.

The Austria-Forum is not organized as one alphabetic encyclopedia, but rather as a collection dealing with topics in different categories, those categories often subdivided further, i.e. Austria-Forum has a more or less hierarchical structure of categories and contributions. The “more or less” comes from the fact that the same contribution can be element of a number of categories. However, this is generally not done by physically copying contributions but by transcluding them or part of them, much as described in Ted Nelson’s pioneering work [3]. We will return to the transclusion concept in detail in later Sections.

The advantage of having categories is that users can narrow their scope of interest before carrying out a search. Thus, if one wants to look up biographies of important persons in Austria one may be well advised to first select the category “Biographies” lest a search for something like “List” also produces as results a number of entries that have nothing to do with persons with name “List”. Each category offers two alternatives: a search just within that category or a search in the whole server. The search can be a full-text search (in this case the restriction to subcategories is of course particularly important) or a search by title and metadata as provided. There are further options...

Manuscript received April 2, 2014. The work was partially supported by the Sparkling Science Program http://austria-forum.org/af/Sparkling_Science.
Hermann Maurer http://www.icim.edu/maurer is Professor at the Institute of Information Systems and Computer Media, Graz University of Technology, and Member of the Board of Academia Europea http://www.ac-info.org.
like allowing “fuzzy” searches ignoring misspellings, etc.

It is important to understand the intention of meta-data: ideally, meta-data should follow a well-defined ontology. This is only partially implemented right now, but it is the introduction of categories that makes it a bit easier than usual: after all, the ontology for plants or persons or cities will be rather different. Thus, a search within one category can be restricted to a much smaller ontology.

Austria-Forum is based on a JSP-Wiki System, enriched with many plug-ins. It contains Wiki-pages as one would expect, but it also includes so-called Web Books: full books with page-turning, full text search, zooming, and other features. Most important is the fact that a link can lead from an entry in a Wiki-page to a different Wiki-page, but can also lead to a page of a Web Book, and conversely. Thus, Austria-Forum is tying together ordinary Wiki material with a selection of (fact and culture) books in a rather novel way. We will return to further plans on how to weave Wiki-pages and pages of Web Books together in a later section.

To get an idea of the size of Austria-Forum it has (end of June 2014) some 500,000 objects, an object being a piece of text, a picture, an audio- or video-clip or some other media object. It is run by Graz University of technology, currently supported by 6 other major Austrian universities, three governmental agencies and three companies. It is also supported by some 50 sponsors [4] and a scientific advisory board [5]. There is a team of 2 programmers and some editorial staff at Graz University of Technology. The team is led by 4 “editors in chief”: the managing editor in chief is the author of this paper. Editorial work is supported by over hundred members of an editorial board [6]. Austria-Forum is surviving without any advertisements.

Despite the sizeable group involved one may wonder how it was possible to compile as much material as is already present, and how further substantial expansion is possible. There are indeed five reasons for this:

(i) It was built on a general Austrian Encyclopedia [8];
(ii) It has relied heavily on material in books, or incorporated full books as Web Books;
(iii) It has incorporated existing data-bases of pictures, see e.g. [12];
(iv) A number of members of the editorial board have built-up substantial collections of information in their area of expertise;
(v) It has allowed incorporating material from other servers (where this was possible due to copyright issues), in the process “freezing” and “verifying” the information by experts whose name, their CVs, and comments added by them guarantee a good level of quality.

2. WHY IS AUSTRIA-FORUM IMPORTANT

The main credo is: We need verified knowledge rather than information of unknown quality. We need it for schools, universities, studies, well-founded decision on all levels, discussions, and everyday life.

The world is basically on an “information-drip” of search engines and servers with information of unknown quality like Wikipedia. Search engines give results ranked by unknown algorithms, and users end up on servers whose reliability is unknown.

Wikipedia is a tool used by “almost everyone”. It is impressive with its many detailed contributions and as good starting point, particularly due to the list of references usually included. However it suffers from a number of facts:

(i) Over 40% of contributions have a personal bias [7]. This in itself is not surprising: anyone writing a contribution is likely to have some preferences. It is dangerous however, if authors are not known, hence one cannot take a possible bias into account. Unfortunately:
(ii) Almost all contributions are anonymous, compounding the problem mentioned
(iii) Anonymity has another unfortunate consequence: some authors allow themselves to be a bit sloppy, since what they write cannot easily be traced to them and “the community will take care if there are some small errors.” Ardent Wikipedians will not like this statement, but the author has positive proof. He identified persons behind some pennames and asked them to publish the contributions under their real name: they refused, explaining that in this case they have to check some details more carefully!
(iv) Wikipedia is often quoted as a good example of the “Wisdom of Crowds” [10] in that some contributions have many authors, or a main author with others adding or correcting information. This has indeed worked well in many cases but it has also lead to the withdrawal of specialists who write a carefully composed essay and then suddenly find it modified and distorted by others.
(v) The author considers it a weakness of most encyclopedic collections (be it in print or be it in electronic form) that for one topic there is usually only a single entry. Often, different views presented in a pointed fashion (rather than attempting to find a “wishy-washy” compromise) would be more illuminating. Also, another phenomenon is emerging: the same topic may need different presentations for different readers:
longer and shorter, more technical or less technical, etc.

(vi) A reader of a Wikipedia contribution has no guarantee whether it is objective, complete and whether what it says is still applicable. Again, anonymity makes this worse, since the reader cannot contact the author for comments or clarification. The issue of objectivity has been dealt with when discussing bias. The issue of completeness is overlooked in its importance: many contributions in Wikipedia are correct, but incomplete. Incompleteness is very dangerous. A simple minded example will explain the problem. The statement „Austria has three mountains higher than 10,000 feet” is absolutely correct. Austria has many more mountains higher than 10,000 feet and but somehow the wording suggests that there are only three. Many more subtle and hence particularly dangerous examples of this kind can be found in Wikipedia. The last point mentioned (whether some expositions are still applicable) is made dangerous by the fact that a contribution may indicate that it was modified recently (suggesting that it is up to date), yet only, say, a punctuation mark was changed. Thus, the fine granularity of the version control in Wikipedia is not always an advantage. Again, the author is aware of concrete examples: e.g. of a contribution on some aspect of computers written in 2006 and even included in the “excellent category” of contributions, yet today the contribution is totally misleading if not read as historic essay … and readers are not warned about this.

(vi) There is clearly some (mild?) kind of censorship of contributions. Again this is made worse by the fact that neither authors nor censors are known and usually cannot be contacted.

Not surprisingly, Austria-Forum is trying to avoid the problems mentioned. The most important point is that contributions are supposed to have an author or equivalent whose CV is known and who can be contacted via a “comment” mechanism: this mechanism allows a conversation between users and authors without either of them giving away their emails, important in an era of rampant SPAM.

Much information in the Austria-Forum is based on books or picture archives. Thus, readers know the publisher, the editors, the authors, and the creation dates of material involved, hence have a feeling about any possible bias and are aware of the date when contributions were considered to be valid. Note that the contents of a book as such are “taboo”, not changes are allowed within pages, yet comments can be added at them bottom of any page, visible as such.

An attempt is also made to assure that contributions can be cited. To be citable implies that the contribution does not change any more and has a well-known source. Note that this “stability in time” seems to run counter to a main advantage of electronic publications that may be expected to always be up-to-date. The compromise taken in Austria-Forum is that at some stage a contribution is “frozen”, i.e. cannot be changed any more, yet comments can still be added to the contribution, hence “the wisdom of crowds” can be used to update information (through comments). Once a contribution is completely outdated (it is a historic document at that point) a new contribution (with links to earlier versions) can be added.

Thus, the Austria-Forum not only allows many contributions on the same topic from different points of view, but also from different points in time. Since it incorporates general and special purpose encyclopedias over the last 200 years it has also become a tool for historic research, for research in how society and language changes, etc. To be specific, it contains a small encyclopedia from 1955 (one version for adults, one for children). It is amazing to see not just the change of contents (like dinosaurs still being thought to be purely vegetarians), but also to note the change of language, and the change of role models. To give one concrete sample, in the books at issue any picture showing a kitchen shows a woman with three girls, any picture of a repair shop a man with three boys.

One major source of information comes from importing information from other sources, if these permit such import. Import is only done by “freezing” the material and by experts identified by name and CV. Since experts will not want to ruin their reputation by verifying material that is doubtful, only material of acceptable quality is imported. Further, in the process of importing, experts can add correction, comments, other sources, doubts about some part of the contribution, etc.: yet they never meddle with the contribution as such, but all changes are recorded under their name clearly distinguishable from the original document.

Note that verification provides also the incentive that the verifier’s name and (if wanted) the institution of the verifier is shown, hence “advertising” with pride the person and institution as contributing to the community. More on verification is found in [11] in this very issue of this Journal of Internet Research.

3. FEATURES AND PROBLEMS

Austria-Forum is work in progress. It is based on a JSP Wiki system and enriched by many
plug-ins, especially an e-Book reader for “Web Books”.

Web Books and how they are used in conjunction with ordinary Wiki pages are certainly one of the “unique selling points” of Austria-Forum. They allow links from book pages to other book pages and Wiki pages, and of course links are also possible from Wiki pages to other Wiki pages or Web Book pages. However, to make full use of this feature it is necessary to be able to create links automatically. For this purpose two developments are necessary: it has to be possible to add metadata to every book page, so that Wiki pages and Web Book pages can be linked automatically; this is the easy part: the more difficult part is to extract from each Web Book page meta data relevant for that page.

This issue is one of the core issues addressed in the Ph.D. thesis [13] and can only be discussed here very superficially. If the book is a kind of encyclopedia with items sorted alphabetically things are “easy”: one associates with each page the topics explained on that page; if the book contains a register then it can be used in the obvious way to associate keywords (meta-data) with book pages; if a book has a list of illustrations, the name of the illustration can be added as meta-data to the book page. Similarly, a detailed table of contents can help. In the absence of all this complex heuristics have to be employed to associate suitable keywords with the pages of a Web Book. In summary, it appears that extraction of meta-data depends very much on the book involved: no general algorithm is possible. All this is further complicated since some old fonts cannot be recognized well by OCR, and that both language and also names of geographical entities change. What once was “Agram” is now “Zagreb”... and the area of the former Austrian empire is of course full of such cases.

Note that other servers like [14] (discussed also in this issue of the journal) are faced with the still bigger challenge to handle old (handwritten) documents.

Austria-Forum has a number of special plugins to add geographic information, to ease searching, to achieve easily nice layouts, etc. etc. Two features warrant special mention: one is a plugin for displaying in a nice fashion a large collection of pictures, applied in many places: in [15] to some 3.500 etchings that are over 100 years old, in [16] to some 2.000 flowers to mention two instances. The other plugin is based on the transclusion idea of Nelson [3]. By specifying in a Wiki page x at point p the code "{{{InsertPage page="URL"}}}" the complete contents of the page with the URL mentioned is displayed (at runtime) at point p in x. This allows the incorporation of a material at some other point without physical copying it. It is also possible to just include parts of a page by specifying a section. However, the section specification may need editing privileges on the page of which a part is to be used: this restriction and the restriction to Wiki pages for this Plugin is a weakness to be removed in the future, see next Section 5.

4. LONGER TERM PLANS

One of the most important aspects that have to be dramatically improved (and indeed work has started on this) are the Web Books. It has been mentioned before that they should turn from page-turning tools as one finds e.g. in [17] and more complex derivatives in [18] into sophisticated tools providing important functions.

One is the “weaving together” of Wiki pages and Web book pages mentioned earlier. Another one is to be able to add comments and links, to highlight paragraphs, etc. for oneself or a designated group. This of course requires sophisticated administration of user groups, also of interest in other contexts, like allowing a class in a school to work on material visible only to that class, at least while “under construction”. A third one is to allow the community to help in correcting texts that have been created by OCR. Note that books when scanned are available first as images only. To make them searchable (and to make extraction of meta-data possible) they have to be converted to text, usually using OCR. However, OCR is not perfect, particularly not in case of some old fonts. A typical case is the famous 60 volume “Wurzbach” that not only uses old fonts but where letters (!) rather than words are spaced unevenly to obtain lines of text of equal length. This, and the variation in font size makes OCR rather tricky, but using learning software and some clever measuring the aim to make Wurzbach full-text searchable is slowly turning into reality, see [19].

A further extension that is not technically trivial is to allow transclusions of any material (be it part of Wiki pages or pages of Web Books) without needing editing privileges for the material of which part is to be used. Note that this touches also on a number of legal issues: transclusion should usually take the form of a citation, so that it is clear where the information comes from. But whether this is shown automatically or only by following a “hint” will depend on the circumstances. Sometimes transclusions may not be allowed for copyright reasons and the system then should not provide this function for that particular document. Observe in passing, that Nelson [3] also proposed transclusion as a mechanism for fair
payment for information. With the current hype for open data the idea that creators of valuable information should also be financially rewarded is largely forgotten, leading to a possible loss of quality as pointed out in detail in [20], hence Nelson’s ideas may witness a revival at some time in the future.

One main issue for any server today is to ensure good access, independent of the user’s device, i.e. independent of whether a PC, a tablet, or a smartphone is used. To assure good readability on all devices is a complex matter, made even more complex if also visual challenged persons should be able to access most information.

A major challenge all information servers are facing is to allow searches using well-structured meta-data, usually based on ontologies in the field at issue. The problem is not to include searches based on ontologies, but to find or define suitable ontologies and then to generate for every document all metadata necessary. Servers with a long history like the Austria-Forum are clearly at a disadvantage, since when work was started meta-data and other aspects of the semantic web were still in its infancy. Hence a major effort to update metadata for Austria-Forum entries will be unavoidable at some stage. This will also necessitate migrating to powerful database mechanisms to hold all information in contrast to the fairly simple minded rather hierarchically structured information repositories used in most Wiki systems. The problem to further improve searching is not restricted to the proper use of meta-data but involves also the difficult questions of ranking search results, of possibly customizing them for individuals, or by novelty, or by popularity, etc.

A major issue on most information servers is the fragmentation of knowledge. If one wants to find a comprehensive exposition on some topic one usually ends up with some search results that lead to further information, and so on. It would be valuable to have features such as knowledge maps that show all relevant information, and have those maps customizable by zooming in, by showing summaries only, by excluding certain issues, by allowing shifting in time, etc. etc. It is the intention of Austria-Forum to provide what can be called “topic lists” as first step, like “Electricity for grade nine in high school” and such. One specific effort to be incorporated into Austria-Forum is a large collection called “Geography of the World”. A glimpse of first attempts to visualize certain geographic facts is explained in [21].

As has been mentioned earlier, Austria-Forum will have an increasing number of “frozen” contributions to allow solid citations. Those contributions will be kept in an archive, but also different views on the same topic are supported. To be concrete, the mechanism envisaged works along the following lines. Suppose there is a contribution for topic x, and a new version y to be added. In this case the editing menu on the page x has the option: “Freeze and add new version” and “Add different point of view”. In the first case, x is frozen (if not already the case), y is added as new contribution with a link to x, but searches will henceforth yield y rather than x (unless a time parameter is specified, see below). In the second case, y is added to the topic list x belongs (or if none exists, as topic list is created) indicating that concerning x there are a number of different views.

One of the most important changes of the Austria-Forum will be the increasing use of interactivity. Thus, Austria-Forum will turn into a repository of information that can be experimented with in many ways as is becoming more and more common with the use of Apps. In order make Apps universally useful, it makes sense to have them run on the server, rather than as an add-on to the device (as is often the case in proprietary software for e.g. the i-Phone). The best current example in Austria-Forum is [22]. 295 students of 5 schools wrote essays on a range of topics. It is possible with for instance [23] to compare the choice of topic, by school, by gender, by age, etc. It is this kind of experimenting with data that will also soon be available for population figures of large cities in Austria, for biographies, for monuments in Austria, for identifying flowers, and much more. Austria-Forum, like large parts of WWW, is starting to turn from a search or menu based information server into a data-base of information open to uncountable experiments.

This tendency will be further strengthened by the addition of new media objects, 3 D scans of artefacts whose very strength is the possibility to manipulate and examine them in a variety of ways.

5. CONCLUSION

In this paper we have described a major effort in establishing an information server Austria-forum.org with quality controlled information, including the weaving together of webpages and pages of electronic books, Web Books. Efforts to improve the system both from a technical and contents point of view will continue, with interactive components gaining importance as one part of the long-range perspective.
REFERENCES

[17] https://openlibrary.org
[18] https://openlibrary.org/developers/api
Web Books:  
The Fusion of Paper and Pixels

Johanna Pirker; Gerhard Wurzinger; Heimo Müller

Abstract - Digital and online reading is now an everyday possibility and is supported by different services and devices. However, the digitalization of traditional books supports even more innovative forms of reading and experiencing the content. In this paper we introduce our solution of an interactive web book library which allows experiencing old books in an innovative and social way. First, we analyze and compare the strengths and weaknesses of existing electronic solutions, and discuss implications for adoption and digitalization of books. We will then present a digital library specialized in old books, which integrates different forms of interactivities and reading experiences. In this context we discuss the potentials of an interactive web book representation using the example of a unique Austrian encyclopedia. We conclude with a discussion of how future reading can revolutionize the way we are experiencing traditional and old books.

Index Terms: E-Books, Web books, Internet books, Digital library, Interactive library.

1. INTRODUCTION

Online reading became an important mean of daily technology interactions. E-readers, reading devices, and digital web libraries with integrated online book readers are designed to simulate the traditional reading experience. Such technologies are being used to facilitate, enhance, and even replace traditional reading. More and more people read texts online and virtually. Especially devices such as Amazon’s Kindle or Apple’s iPad have revolutionized the way people experience literature and other readings [10]. New forms of reading also changed the reading behavior and the user’s expectations towards the digital content. They expect realistic behavior and the look and feel of traditional books combined with interactive features such as searching, highlighting, or bookmarking.

However, still many users have a negative attitude towards technology-enhanced reading of books. As we will explain later in this article, attracting users to these new forms of reading require advanced designs of interactivities to show advantages of digital reading compared to traditional reading in a familiar setting. Also, different kinds of applications require distinct tools and possibilities to interact with the content. For the different purposes such as academic research, learning at school, or leisure, the required tools change.

In this article, we shall discuss different interactive and social technologies to enhance the traditional reading experience, shifting the focus of design away from simply simulating traditional books to enhancing reading experience with interactive and social technologies. This should not only support the new forms of reading, but should also help attracting people to read more books online and to support different application domains such as academia or research.

Our goal is to study different techniques to improve the technology-enhanced reading experience. The contributions of this work include, first, a characterization of social and activity-based interaction styles to enhance the reading experience. Second, we introduce web books as integrative solution based on these characterizations and introduce this concept in a case study. We close by discussing future advancements and ideas of web books.

2. BACKGROUND

Research in different areas is relevant to this work: (1) digital reading technologies and (2) online libraries. We consider each in turn.

2.1 Digital and Online Reading Technologies

One of the most popular forms of digital reading is the use of electronic books (e-books). E-books are usually formatted specifically for a reading device or software, and can be either created manually, or can be a result of a digitization process [1][9]. In contrast, digitized books are virtualized (e.g. scanned or photographed) books, which are usually stored in high-resolution formats such as TIFF and can be further processed with optical character recognition (OCR).

Manuscript received April 2, 2014.  
J. Pirker is with the Institute for Information Systems and Computer Media, Graz University of Technology, Austria (e-mail: jpirker@iicm.edu).  
G. Wurzinger is with the Institute for Information Systems and Computer Media, Graz University of Technology, Austria (e-mail: gwurz@iicm.edu).  
H. Müller is with the Medical University of Graz, Austria (e-mail: heimo.mueller@mac.com).
recognition (OCR) algorithms [4] to be readable by machines. The OCR content can be the basis for the creation of e-books or other forms of digital representations.

Supporting machine readable content, the text could be presented in several innovative ways. However, most digital book readers are designed to present the digital content in a natural way. Studies suggest that users like to keep functionalities they used in traditional paper books also in the digital representations [4]. Hence, many digital readers assimilate the characteristics of real books, and integrate functionalities and visualizations such as page turning or inserting bookmarks, what should increase the usability and navigability of the medium [3]. In particular reading for pleasure focuses on the readers’ satisfaction and enjoyment and should remind users of traditional reading experiences. For this purpose, Wilson, Landoni, and Gibb [16] suggest the integration of book design elements which meet the expectations of readers based on their experience with printed books. This includes visual elements (e.g. book cover), elements to support a sense of structure (e.g. table of content), and a sense of place (e.g. bookmarks, progress bars).

The shift from traditional to electronic books also results in the integration of different interactive functionalities and enhanced reading experiences. Modern e-books support different advanced interactive features. Apple’s iBooks, for example, supports widgets such as image galleries, audio and video players, review questions, 3D-object viewers, and HTML objects [10]. The integration of multimedia and interactive elements can have a positive impact on the readers’ engagement with the book. This may also enhance their ability to understand and remember the information [16].

Furthermore, such interactive possibilities allow the use and application of reading far beyond traditional reading. Digital reading is an important topic in areas such as leisure, learning, and research. Different school formats already integrate digital reading models into the classroom. Some colleges already use e-book readers such as Amazon’s Kindle instead of printed books and also many libraries support digital formats [19]. Fenwick et al. [10] describe the application of interactive digital books as an interactive textbook experience with Moodle integration. In learning scenarios and research it is particularly important to focus on comprehensibility, searchability, and content linking. [15] found that it is important that the digital reading device supports responsive reading techniques and the possibility of generating new texts to support reading in academia.

2.2 Digital Libraries

In online libraries and digitalization projects such as the Google Books Project, the Open Library Initiative, the Internet Archive, or the Million Book Projects a large collection of digital books can be retrieved. Their main purpose is to collect, preserve and organize books in digital form, and make them searchable and retrievable [14].

Many of these projects are converting traditional books into digital representations performing OCR on an industrial scale [5]. This mass digitalization allows users to retrieve a wide range of different books. However, the output is far from interactive content and is mainly used for preservation and discovery [6].

Another issue of traditional digital libraries is the information integration of other sources. How well digital books and digital libraries can cooperate with other information sources is an open question. There is only little work on linked libraries. In this paper we address this gap, offering a web book library solution, which is integrated into the online encyclopedia and media content collection Austria-Forum.

3. INTERACTIVE AND LINKED WEB BOOK LIBRARY

We have seen different aspects of interactive books and libraries. In this section, we want to introduce our solution of an integrated digital library, which is a collection of interactive and integrated digitized books. Through the strong connection and integration of the web book library with the Austrian online encyclopedia and information portal Austria-Forum [2][8] a new form of reading, researching, and learning is possible.

Web books (also Interactive Internet Books [9]) are interactive and social digitized books supporting reading and searching in both, in the original high-quality scans, or in a cleaned OCR-processed content. Similar to other digital readers, our web book reader assimilates the design of real books.

Figure 1 shows the selection of a book in a bookshelf.

Figure 2 displays the reading experience in a design simulating a printed book. Adding interactive content and new features should enhance this design, but should not distract the user from main purpose, the reading experience. Hence, new forms of interactions are added as layers.
Figure 1 Book-shelf representation in the Austria-Forum web book library

Müller and Maurer [9] describe the structure of web books as a layered structure. Each book consists of four layers: (1) the facsimile layer is used to display the original version of the book. High resolution images give the users the feeling of reading the original. (2) The second layer, the OCR (optical character recognition) / text layer, is the computer-readable text representation of the scanned text content. (3) The enhancement layer allows additional annotations such as personal markers and notes, links, highlights, and nano-publications (“a set of annotations that refer to the same statement and contains a minimum set of (community) agreed upon annotations” [7]). (4) The communication layer is responsible for social interactions such as content sharing, social tagging, discussions, reading history visualizations, and search agents.

Figure 2 Reading in an interactive web book

This layered structure with some simple interactivities and the connection to the Austria-Forum of our web book library offers the following main features:

(i) Bidirectional behavior of Information Integration.

While most internet books only link into one direction, the integration into the encyclopedia Austria-Forum allows several further features, such as bidirectional linking. Knowledge content from Austria-Forum can be enhanced by web book content and web books can be enriched by Austria-Forum content. This is supported by the two major principles links and transclusions.

Links represent references to other sources. The web book reader does not only support the link to internal and external URLs, but also to different media types, such as images, audio, or videos. (See Figure 3)

Ted Nelson’s idea of transclusion (inserting and embedding the original content through shared instancing) [13] is applicable to the library. While the encyclopedia pages already support the inclusion of referenced pages (using ‘insert page’ commands), the same principle should be possible with book content. Transclusion has different advantages compared to linking. Users have the possibility of exploring the original content instead of looking for the context in linked pages, the linked content always uses the original and updated information, and less disk space is required [11]. More on the integration of transclusion in the context of Austria-Forum can be found in [8] in this issue of this Journal.

Figure 3 Linking content and media into the web book

(ii) Interactive and Engaging

The web book library supports different interactive functionalities. Users can highlight content, add comments, add bookmarks, or share content. Switching between the original facsimile representation and an OCR-processes text version makes even historic books available to all user groups (e.g. without knowledge how to read fraktur). Users can view the content such as images in original or zoomed size. The search functionality of the book gives the users an overview of the search results on each relevant page. This makes it easy to find context sensitive content (see Figure 4). Not only the results on the page {1}, but also in the entire book {2, 3} are displayed.
(iii) Social

Other important features which have become more and more relevant in the last years are social and community-based interactions. Users don't only want to read content, they also want to share and discuss content immediately within a social network. The web books support group-intern social interactions, which means that marked, shared, or commented content is only visible to a particular working group. This allows different forms of online collaboration (e.g. collaborative research), community projects (e.g. the manual correction of OCR content), or collaborative e-learning (e.g. team projects for school classes).

Most of the introduced features are already realized and integrated into our web book solution. Features such as transclusion and group sharing are still under development. In the next chapter we describe the process of converting historic printed books into web books.

4. Integrating Historic Books into the Web Book Library

The web book library of the Austria-Form is in particular famous for its representation of historic books. In this chapter we discuss the single steps necessary to integrate a (historic) book into our web book library. First, the original book must be scanned and digitalized using OCR software such as Tesseract or Abbyy Finereader.

Digitalizing old books to be representable as an interactive web book, however, brings different implications. Far beyond the implications of optical character recognition of traditional books (e.g. scanning defects), old books often require special attention regarding their font style (often fraktur), their general condition (completeness, color, or damages), and the dictionary used.

An especially interesting case is the biographic 60 volume encyclopedia ‘Biographisches Lexikon des Kaiserthums Oesterreich’ by Constantin von Wurzbach (also referred to as ‘Wurzbach’) [18]. The original scans pose several challenges for OCR software and are especially hard to process with traditional tools without further modifications and enhancements. In particular different Fraktur typefaces, frequently changing font sizes and styles, arbitrary character spacing, the use of different languages (German, Latin, and Italian), and obsolete spelling and vocabulary turn Wurzbach into a challenge for common OCR algorithm (see Figure 5).

For this special case traditional OCR algorithm fail to recognize the entire text in an adequate quality (see Figure 6). This level of quality would neither allow acceptable reading of the text, nor searching for specific content.

For this special case traditional OCR algorithm fail to recognize the entire text in an adequate quality (see Figure 6). This level of quality would neither allow acceptable reading of the text, nor searching for specific content.

Wikisource [17] provides manually created and corrected data of the first volume out of 60. To enhance the OCR algorithm we have created an own dictionary based on this data. This dictionary of the used vocabulary can be generalized to be used as a basis to train the dataset for the remaining volumes. We use the manually created volume to control the quality of our algorithm. Different quality measurements such as Levenshtein distance or text analysis are used to
compare the machine-processed text with the manually corrected one, to be able to optimize the algorithm for the remaining 59 volume.

Wurzbach is a unique collection of biographies of persons (in particular of the Habsburg monarchy) living between 1750 and 1850 in the Austrian crown lands. Integrating all volumes in a machine readable way into the Austria-forum web book library and connect it to the Austria-Forum content will allow users to search, research, and connect persons.

5. Future of Web Books

Our future research will focus on accessibility and interaction with knowledge embedded in web books. For this we plan to develop algorithms for automatic knowledge discovery and data mining (KDD) optimized for (encyclopedic) book collections. Such algorithms provide functionality for rule mining, subgroup discovery, graph mining, queries design, and the analysis of structured, cross-domain data sets, and in consequence allow identifying valid, novel, potentially useful, and meaningful patterns within a web book, which can be accessed as linked open data, e.g. in the nano-publication format [7].

Future user interfaces for knowledge discovery in web books will build on well-known principles of communication and social interaction in the support of comments, links, and semantic tagging of statements in nano-publications. For the manipulation of “facts”, we will develop a special editor, which will on the fly generate semantic tagging of statements in the OCR text using software agents, that can deal with open, rich and ambivalent ‘natural’ inputs, such as digitized books and engage the reader to provide their knowledge to train agents or share knowledge directly with other users. Such interactive agents can semi-automatically identify meaningful nano-publications in web books and consolidate them using the knowledge provided by readers of a web book. Finally a web books and its embedded itself becomes the interface, and embodies the link between actors (readers of a books) and concepts, which are described in the formal representation by RDF triplets. By the behavior of a large number of users (who, when, and which parts are used in a certain context) we can build a "social ontology" of the web books, and extract facts as open data which can form the base for a central knowledge repository similar Wikidata or Freebase.

6. Conclusion

In this article we discussed different techniques to enhance online reading experiences with focus on integrated, interactivity, and social experiences. We introduced this concept as interactive web books based on the example of a digital library, which is integrated into the online encyclopedia Austria-Forum. This article’s contribution is a model to create book experiences, which are interactive and socially attractive and integrated into a knowledge database to support not only leisure reading experiences, but also innovative and intelligent research and learning.

In future work we include research on automatic knowledge discovery and data mining algorithms for book collections and will integrate concepts such as nano-publications and will introduce user interfaces with even more social and interactive behavior. Furthermore, we focus on the integration of an advanced transclusion system, which does not only allow the integration of books pages into encyclopedia pages, but also additional inserted content into the web books.

Acknowledgments

We would like to thank Hermann Maurer and the members of the Austria-Forum team. The Wurzbach project is done in collaboration with the University of Vienna (Thomas Winkelbauer), and has received much help from the digitization group of the University of Graz (Karl Lenger).

References

Maurer, H., Müller, H., "Can the Web turn into a digital library?,” International Journal on Digital Libraries 13(2), March 2013, pp. 65-75.


Wurzbach, C. v., „Biographisches Lexikon des Kaiserthums Oesterreich,” 1891.


 AUTHORS

Johanna Pirker studied software engineering at Graz University of Technology and concluded with a thesis on virtual worlds for physics education in a joint project with Massachusetts Institute of Technology. Currently she is a researcher at the Institute of Information Systems and Computer Media, Graz University of Technology.

Gerhard Wurzinger is currently a software engineer and researcher at the Institute for Information Systems and Computer Media, Graz University of Technology. After some years of professional experience as software engineer in the automotive industry Gerhard Wurzinger received a degree in Software Development and Business Management at Graz University of Technology. He is the technical lead of the Austria-Forum.

Heimo Müller studied mathematics in Graz and Vienna, concluding with a thesis on data space semantics. As Marie Curie fellow, Heimo Müller worked at the Faculty of Arts of the Vrije Universiteit Amsterdam in an interdisciplinary program on word and image studies and was the founding head of the Information-Design programme of the University of Applied Sciences FH Joanneum in Graz. At the Medical University of Graz he developed an interactive data exploration system and investigated large inhomogeneous clinical data collections and worked on the ICT architecture of the FET flagship proposal IT Future of Medicine (ITFoM).
Verifying Information on the Internet

Korica-Pehserl, Petra

Abstract - The Web and its importance is steadily growing. However, how much of the web data is stored on reliable servers? How much of the information on the Web and especially in online digital libraries like Wikipedia can be trusted? In this paper we propose the process of verification for web data and we describe how this process is implemented in Austria-Forum. Furthermore, we concentrate on general aspects, benefits, and potential pitfalls of verified knowledge.

Index Terms: Web, Digital libraries, Wikipedia, Information retrieval

1. INTRODUCTION

The Web is growing at an incredible speed! For example, the latest statistics say that alone in Austria there are dozens of new servers each day. But how much of this web data is stored on reliable servers (e.g., where it can’t be manipulated without a trace, or where the information is biased on purpose or due to lack of expertise)? How much of this data appears in results of search machines? To what extent is this data valuable and reliable? We argue that only a very small amount of information on the Web servers is reliable and quotable. To be quotable, a piece of information must be stable over time and have a reliable and verifiable source.

This is especially crucial in education because we want our children to learn from reliable sources and, especially in higher education and universities, we expect our scientists to take only reliable information into consideration.

The author is involved in Austria-Forum [2], an online platform with a goal to serve as a hub for online quotable knowledge as described above. Austria-Forum is a kind of online encyclopedia and a rapidly growing collection of “Austriaca”, i.e. documents that either have to do with Austria (or with the region interpreted broadly since Austria was much larger before the end of WWI), or have to do with Austrians, or express something of value for Austrians. However, Austria-Forum is not one encyclopedia, but a large set of some 50 specialized encyclopedias, in addition to a general one. Material is stored and presented in the form of text, pictures, video and audio clips, and other multimedia elements.

One major source of information in Austria-Forum comes from importing information from other sources, if, that is, these permit such import. In order to realize the import of relevant data, Austria-Forum implemented a verification process which we describe in detail in this paper.

2. VERIFICATION AND DEDICATED SERVERS

As described in [3] verification is a process that “proves” that an article is correct in the current version. In verification the current version of an article is imported into Austria-Forum, gets “frozen”, and is verified by an expert or community of experts or by stating a reliable source, like a verified expert or verified print media, as “proof” of correctness. Since experts will not want to jeopardize their reputation by verifying material that is doubtful, only material of acceptable quality is imported. In the process of importing, experts can add corrections, comments, other sources, as well as express their doubts about some part of the contribution: yet, they never change the contribution as such, and all changes are recorded under their name clearly distinguishable from the original document. After verification, new information can be added using a comments field or by adding a link to new information on that article, but the originally verified article doesn’t change anymore. There can be exceptions in the process of verification, as, for instance, it does not make any sense to verify articles dealing with current topics and facts that are likely to change significantly in the near future.

This is not a problem because, in distinction to other online encyclopedias, Austria-Forum has its own criteria regarding the time concept of the articles. Most of the articles deal with topics that are time-proof like history, biographies, places, natural phenomena ... The half-life of most items in Austria-Forum is considerably longer then in Wikipedia, and it is exactly these types of articles that are especially targeted for verification because they describe topics which don’t change much over time and which are important for general discussions, for awareness and for education. A concrete example is the integration into school curricula that Austria-Forum is involved in at the time of writing of this paper. Pupils use the Web to learn how to perform
scientific work. The scenarios are multifld, especially when we add the usage of tablets in classrooms.

Austria-Forum is currently supporting a school initiative from the Austrian Ministry of Science and Research called Sparkling Science Program [5], with the goal to teach students aged 11-17 how to work in a scientific way. Teaching is an example but not the only task where one would profit from reliable and quotable knowledge on the Web. We believe that we need, in addition to Wikipedia and similar Web sites, other specialized dedicated servers with emphasis on local history, achievements, features of nature, personalities, cities and villages, culture, etc. Furthermore, such servers should concentrate on items that do not become obsolete quickly, and should also develop new approaches to searching and preservation. Austria-Forum is using some information from Wikipedia based on the creative common license to enrich Austrian digital heritage, and verification of such Wikipedia entries is an important part of this project. Other sources of openly useable material will also be used in the future to enrich existing material. There are many other potential sources with data from different domains, for example CIA World Factbook [4], list of geographical information [12], [13], list of recipes [14] among others. Of course, in case of different materials that need to be verified, there might be a need for different algorithms for extracting and verifying information as the structure of the source is different.

Regarding Wikipedia we encounter the issue that quite a few of the articles deal with a topic in an incomplete manner or only deal with one aspect of the topic, and the content of search results is not prepared to give the user a fast overview of the topic with all the pros, cons, trending, experts etc... For example, the German Wikipedia article about „Fettembolie“ (Fat embolism) see [1] was written by a doctor specializing in orthopedics, and the article only deals with fat embolism from this viewpoint. Omitting other cases of the particularly dangerous fat embolisms that can occur in the brain and the lung can cause the reader to misunderstand, misjudge, or misinterpret some important facts. Often articles in Austria-Forum also go deeper, in terms of knowledge and information provided, than the Wikipedia ones. One of the goals of Austria-Forum is to describe a topic from various perspectives in order to grasp the full extent of the complexity repercussions of a topic. If we take climate change as an example, the reader will find various articles on that topic in both Wikipedia [9] and in Austria-Forum [10]; however, what is missing is a sort of verification of the results and ranking various search results based on their topics; e.g. pro or contra in this case, showing trends of the results, related search queries etc. in order to give reader all necessary information as objective as possible.

3. APPROACH - HOW DOES VERIFICATION WORK

As described in the previous chapter, verification works by freezing an existing article and proving its correctness via another reliable source or by an expert.

3.1 Benefits for verifiers

Becoming a verifier in Austria-Forum can be interesting for several reasons. Let us briefly discuss the benefits for persons on one hand and for institutions on the other.

For any person, becoming a verifier is a good opportunity to position him- or herself as an expert on a topic. For example, if a person is an expert in the area of, say, „electricity“, he or she could verify articles on that topic. The obvious benefit here is that when people read the relevant articles they are going to see the person’s name at the end of the article and will therefore associate the person with the topic at issue. A large number of verified articles increase this effect. Also, next to the verifier’s name there is also a link to his or her CV which any interested reader can click on to get an impression of the qualification of the verifier.

An additional benefit for verifiers is the recently planned functionality called „topic list“ in Austria-Forum. Topic lists will be semi-automatically generated from all verified articles on one topic. When a user searches for a certain topic in Austria-Forum, the topic list for that topic is displayed as a subcategory next to the search results. For verifiers this means that their name is going to appear in search results for the topics they are experts on, e.g., „electricity“. Additionally, a related topics list could be displayed, providing even more visibility for the verifier. This is especially interesting for teachers, professors, retirees, idealists and other people who are experts on a topic and want to share their knowledge and get recognition for being an expert.

As for institutions, verifying content is also a good opportunity to position the institution as an expert body on a certain topic – for example a computer science school that verifies most of the articles on computer science will be seen as a quality institution as well as an expert body on this topic, which will likely boost enrollment and help the school secure the necessary funding and hence its very future. Of course, the same logic may be applied to companies, and Austria-Forum is currently in discussion with several of those. Some research-related institutions are even considering the act of verifying an article as a
form of publication. This is an interesting opportunity to increase the number of publications which we know is an important criterion for rating of such an institution. It is also possible that museums and similar institutions might start verifying content from their field of expertise – e.g., articles about Austrian art, music, etc.

Upon reading a verified article a user can choose to click – here the quality of the article and perhaps some useful comments and additional information in the verification ribbon will play an important role – on the name of the institution and could be led to the website of the organization. This is exactly what, for instance, a University in Kobe [11] has agreed to do. They are offering Japanese courses for Austrian people, and so they have decided to become a verifier for topics dealing with Japan, thereby positioning themselves as an expert which in turn enables them to have some additional „content advertising“, as users are going to see the institution’s name on the article and are likely to click on the hyperlink leading to its website.

3.2 Quality control for verification

A process ensuring an acceptable level of quality control is necessary. As already mentioned, there is the link to the CV of the verifier next to his or her name. This CV has a double function – to provide an easy way of determining whether the verifier is qualified – the user just needs to click on the name of the verifier; and, secondly, it acts as a form of social protection against users verifying articles they don’t know anything about, in the process even humiliating themselves. If this doesn’t help, there is also a control mechanism on the editorial board level in Austria-Forum that ensures that articles that are falsely verified can get “caught”, in which case the verification label is removed.

In order to become a verifier, a person needs to be first registered with the editorial board. Additionally, a person who wishes to become a verifier needs to go through a selection process where his or her credentials, qualifications to curate a certain topic, and expertise are verified. Upon their registration as a verifier, each verifier chooses one or more fields of expertise.

When designing any verification process there are some crucial questions to be considered. For example: “How do I choose what subset of the verified articles I should control?”; “How do I make sure that the articles I choose to control constitute a representative subset of all verified articles?”; “How do I deal with the ever growing number of articles to be verified?”; or “How do I make sure a newly added verifier is doing quality work?” We propose the implementation of random verification of all verified articles on a monthly basis, every six months, and annually. In addition, the editorial board should verify all articles (if possible) verified by a new verifier in his or her first month of activity, and then expand this process to a random subset of articles for the first six months, taking a small number of verifications to check on a yearly basis.

However, the process described above is slow and limited in scalability. Despite its high quality standards Austria-Forum also aims for scalability and fast verification of many articles due to a high number of materials to be verified. For boosting the scalability we are working on a way to automate the verification process in some cases. Say if we considering n sites on one topic, parsing those sites for features and comparing, then if a feature is the same in all sites, that feature is likely to be true. A concrete example might be verifying CVs of a person - if on e.g. 4 web sites the date and place of birth of a person is the same we may assume that it is fair to say that this information has been indeed seriously verified. Still there are some aspects/questions that need to be considered, for example what is a good number of sites and how do we make sure that those sites are reliable themselves. This kind of verification could be given to an additional verifier class – a verifier who is not an expert him-or herself but possesses written proof of the ability to verify articles.

This could be a reference work on a topic, for example, and the task could be given to a large community of users and could thus simulate a sort of crowdsourcing or an Amazon Turk [7] platform-like effect. Crowdsourcing is the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, especially from the online community [6]. There are various questions to be solved in order to build a healthy community: for example, how to find and attract experts to become part of the community; how to motivate members of the community (e.g., recognition in the community, vouchers, monthly “top verifier” site; …) Additionally, Austria-Forum should consider implementing a class hierarchy system for verifiers, so we suggest that, in order to keep the high quality standards of AF and in the case of articles verified by online community and not by the experts, a small comment be added next to the verification field to indicate that these were not verified by a subject matter expert.

3.3 The process of verification

At the time of the writing of this paper verification is already implemented in Austria-Forum. In order to verify content, any user who has the “verifier” privilege can choose the option “page verification” in the menu on the upper right part of the article to be verified. Then the verification mask is displayed in Figure 1.
Users reading the article can check whether the person ("verifier") is appropriately qualified to verify the present contribution by clicking on the link to a short biography and qualifications of the verifier.

Despite the "freeze" of the verified contribution, additional comments from the community are possible and encouraged. The comments are visible at the end of the article and in front of the verification ribbon. In fact, verification could even be used to bring in different points of view on existing articles.

Once an article has been verified it is currently not possible for this article to get verified again by another verifier. Currently, the only quality criteria is that the article is verified – there is no rating or hierarchy of the articles based on the number of verifiers that have verified them. The notion of having more than one verifier only makes sense in case of the so-called "partial verification". This kind of verification can happen if the article is a longer one and if it deals with various topics. This means that one verifier, who is for example certified for history, will verify one part, while the verifier who is certified for geography will verify another part of the same article. Or another verifier could look at the article but not be able to verify the origin or logic of a formula. He or she could then say "...this article is OK, but I could not verify the formula xy".

3.4 Comments field

As seen in Figure 1, in the current version the information in the comments field is not structured. While this enables the verifier to write and copy-paste all the information he or she wants (including the uploading of pictures or clips and links to other pages) it also makes it more complicated to extract the information automatically. Therefore, we propose to expand the verification mask to implement additional fields and unified separators, e.g. semicolons, between the values with the goal to do information retrieval on the comments field. The comments field is situated in the verification ribbon below any article in Austria-Forum.

<table>
<thead>
<tr>
<th>O fully verified</th>
<th>O partially verified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verifier:</td>
<td>&lt;name&gt;;&lt;name2&gt;</td>
</tr>
<tr>
<td>Timestamp:</td>
<td>&lt;01/01/2014/12:00:00&gt;</td>
</tr>
<tr>
<td>Source:</td>
<td>&lt;source1&gt;;&lt;source2&gt;</td>
</tr>
<tr>
<td>Comment:</td>
<td>&lt;some comment&gt;</td>
</tr>
<tr>
<td>Topic:</td>
<td>&lt;topic1&gt;;&lt;topic2&gt;;</td>
</tr>
<tr>
<td>Keywords:</td>
<td>&lt;keyword1&gt;;&lt;keyword2&gt;</td>
</tr>
<tr>
<td>Additional comments and arguments</td>
<td>&lt;some argument&gt;</td>
</tr>
</tbody>
</table>

Table 1 Comments field to be displayed below an article in Austria-Forum

The verifier inputs data together with references to a written proof - e.g. quotation of a well-known reference work, if applicable, as well as a comment. At this moment the article becomes “frozen”. Through "freezing" (archiving) an article becomes quotable, which is a necessity for all institutions that need quotable general knowledge. Although articles get “frozen” after verification, it is still possible to add additional comments. Furthermore, the link to the latest version of the article (that is maybe not verified yet) is kept up to date.

In our experiment this was done with German Wikipedia because of the relevance of the content and the easy import, thanks to the “Creative Common License”. However, “freezing” contributions does not mean that most information is obsolete, but rather that topical versions are always available without any of the older versions being destroyed. When doing the regular update of Wikipedia entries, new versions of articles are merged into Austria Wiki but the verified articles are not going to be changed.

Once an article is verified, a yellow ribbon can be seen below it under the part “Zertifizierung”, as displayed in the image below. Any verifier can comment on the article. Other relevant information like date, “frozen” quotation link and the name of the verifier are also included. As for links, there remains a link to the current (but then maybe not yet verified) contribution and regular updates are performed, but without changing the frozen posts.

Figure 1 Verification mask in Austria-Forum

Figure 2 Verifier ribbon in the article about the Theory of Relativity
Only approved users can verify and change the information on verification in the comments field; however, any registered Austria-Forum user can use the last part of the comment field to share his or her opinion and knowledge on the topic. As noted in the previous chapter, an additional benefit of the verification is that it will bring in different points of view into existing articles, resulting in a higher degree of knowledge about a topic as well as a healthy discussion base on the same.

The comments field can also be used to implement the functionality whereby more than one person verifies the article. In case of partial verification, the verifier crosses the matching box, writes his or her name, and a comment on which part of the content in question was verified.

3.5 Information Retrieval from Verification

By making the information in the comment field machine readable and searchable, it is possible to use it for data mining and manipulation. For example, all verified data on a topic can be clustered for displaying a topic list related to the search query in the search results. Additionally, this data can be used for generating lists like, e.g., „newly verified articles“ or „these articles were marked as highly recommended by our editorial board“... In addition the so-called “topic lists” can be generated from the titles and verification ribbons of the articles. Those lists should be automatically created by parsing the verified contributions and extracting metadata like authors, verifiers, topics and others. Topic lists like [10] can be used as a form of knowledge clusters and for consolidating information on a given topic. Some of the benefits of topic lists are that they can help identify experts on a topic, find out which topics are currently popular, associate persons and institutions to topics, or find out what the new articles for a given topic are. Last but not least, users should be able to set a recurring search alert on a topic or person/institution and get the results via e-mail on a regular basis.

In order to deal with the big volume and number of articles, topics and verifiers, the metadata can be used for administration purposes, by means of creating sites in a similar way that Wikipedia uses them [8], e.g. “last changes”, “topics to be verified”, „verified sites to be improved“.

4. Conclusion and Future Work

As we stated at the beginning of this paper, we need verified knowledge rather than information of unknown quality. In this paper we have described the process of verification as it is done in Austria-Forum as well as some general aspects, benefits and potential pitfalls of verified knowledge. Verified information is needed to make sure that the information we are getting is “correct” and as objective as possible, and can thus serve as a basis for well-founded decisions on all levels in everyday life. We will continue to improve and expand the verification process and expand the number of sources imported and verified.

acknowledgment

Many thanks to Professor Hermann Maurer and the Austria-Forum team for their help.

References


Authors

Petra Korica-Pehserl Petra Korica-Pehserl studied Software Development and Management at the Graz University of Technology (Austria), from 2003 to 2008, and is currently a Ph.D. Candidate in Computer Science at the same institution. She has worked for Microsoft in Vienna (Austria), Paris (France) and Zagreb (Croatia) since 2009, and is also the author of fifteen papers as well as a member of the Editorial Board of Austria-Forum. So far, Petra has held more than 30 presentations at various conferences worldwide.

Main research and project areas: Web-based learning; languages; artificial intelligence; robotics; computer vision; information retrieval; Web 2.0; semantic Web; data; cloud computing, IOT and applying computer science to business.
Geographic Data Verification
Automatic Verification of Physical Geographic Data Using Maps

Mehmood, Rizwan

Abstract - The possibility to publish online information causes a dramatic increase of data sources on the web, thus reducing the difficulty of finding information. On the other hand, the major obstacle which web users often find is the reliability of information, as some times information differs from one data source to another on the web. Therefore there should be a mechanism to verify online information. We start by defining the term geography and identifying geographic data covering different branches of geography. The prime objective of our efforts is to automatically verify parts of geographic information which is found on the web using different methods and approaches. Our contribution is manifold and the major focus is on two issues; i) addressing the reliability of information ii) verifying geographic data found in different online data sources by applying spatial queries on the world maps.

Index Terms: Geography, Geometry, World countries, Reliability, Information quality

1. INTRODUCTION

There are innumerable definitions of the word geography. Geography is a branch of science that deals with the study of earth and the factors that affect the living things on earth. It is the science of place i.e. the study of the surface of the earth [11]. There are different types of geography and subsequently each type covers a different type of data. We present a classification of geographic data that covers different domains of geography as shown in Figure 1. The major emphasis in this paper is on verifying information related to the physical geographic features as highlighted in Figure 1. For instance, the examples of data that relate to physical geography of any country are: area of the country, neighbors of the country, total length of political boundary of the country etc. One of the key aspects related to geographic information on the web is information quality. The dubious nature of geographic data needs special attention in terms of reliability; for instance the political boundaries of some countries often remain an unsolved issue, Pakistan and India are typical examples of such countries. Where does a river start and where does it end? Are boundaries purely constructional (etc.)?

There are two obvious methods for ascertaining the truth and correctness of physical geographic data i) textual verification and ii) geometric verification particularly in case of physical geography using maps.

Textual verification of geographic data can be done using online encyclopedias such as Britannica and Brockhaus (one of the largest German encyclopedias).

For geometric verification, we have conducted an analytical study that consists of two steps; i) calculation of physical features of the world countries using spatial queries and ii) comparison of results of geometric calculations with the geographic data that is found in different online data sources such as Wikipedia.

The rest of the paper is structured as follows. Section 2 presents a literature review that explores factors contributing towards the reliability of web page and it shows the glimpses of research work conducted in past, to verify geographic information. Spatial data handling tools and techniques that play a vital role in calculation of physical geographic features are discussed in Section 3. Section 4 is based on the actual verification of geographic features using real world examples followed by a conclusion.

2. LITERATURE REVIEW

This section presents an overview of the current approaches and methods that have been proposed for judging the reliability of online information sources. The features contributing towards reliability of an online data source are shown in the Figure 2. Often, information is found on the websites that give no indication of a specific source from which the information is taken. Most of the time, outdated information is presented. Time and date are very important dimensions that make information reliable. A valid indication of the reference, from which information is taken, makes it more trustworthy.

The data is usually considered more reliable if it is found in a website having "gov" domain. The reason behind this lies in the fact that it is

1 http://www.britannica.com/
2 https://tugraz.brockhaus-wissensservice.com/
produced by the government of a particular state, city, or country.

Figure 1: Examples of geographic data covering different branches of geography

Figure 2: Features contributing towards reliability of an online source of information

Wikipedia ensures verifiability of articles by the presence of cited reliable sources so that readers of the article can verify the information using a cited source. The textual verification is a well-known research area in terms of verifying information present on Wikipedia. Calzada and Dekhtzar presented an approach to check the quality of Wikipedia articles using encyclopedias [2]. Weather verification using geographic systems is done by Ken Waters [3]. In contrast to textual verification our efforts are based on geometric verification. This paper is a step towards verifying online data using maps. In order to verify geographic data we have used spatial queries. A spatial query means that we want to select certain data based on its location. Ujaval Gandhi [4] has highlighted the role of spatial queries to verify geographic data. He has presented the verification of geographic data using a case study about populated places located near a few kilo-meters range of rivers. Two shape files that contain data of rivers and populated places in the world were taken for spatial analysis. He used buffer query to calculate the populated places that are a few kilometers from the selected river crossing through a particular region on the earth.

To verify information we simply take data from online data sources particularly from CIA world factbook and Wikipedia. The CIA world factbook is the most widely used online information source for geography. It covers all aspects of geography. This was selected as our starting point. It is worth mentioning here that further geographic information is found in Wikipedia. It presents data in the form of tables and info boxes. The info boxes in Wikipedia contain factual information about countries of the world and other geographic entities such as rivers.

3. SPATIAL DATA ANALYSIS

This section is a brief introduction to spatial data and serves as a background to understand the verification of geographic data in Section 4. Geographic features of earth can be represented by three main types of geometry such as point, line, and polygon as shown in the Figure 3. Point is used to represent any geographic location such as cities, parks, or such. Line is used to represent rivers, roads, railways etc., and polygon is used to represent features like countries, islands etc. There is one more possibility of combining geometries in ‘GeometryCollection’ data type. It may contain any type of geometries. An example of ‘Geometric Collection’ is the country Japan which consists of many islands or “the great wall of china”[1] [8] [9]. There is no exact type of geometry that can be used to represent all geographic features. The choice of choosing a particular geometry to represent data depends on the requirement of the application [1].

Most of the database management systems are providing spatial data handling facilities. This technology enhancement has made it easier for the geographers and researchers to analyze spatial entities. Spatial databases play a vital role in a number of applications such as GIS, AutoCAD etc. as highlighted in [6] [7] [8] [9]. Spatial operators come handy during spatial analysis. For verifying physical features we have used two main spatial data types - namely geography and geometry.

Fig. 3: Examples of objects that are used to conceptualize geographical features like river, roads, and countries

5 Shape file has the extension “.shp” and it contains topological information of geographic features

Geography data type is used to represent geodetic data. Geometry data type is used to represent planar data i.e. data that defines a point on a flat 2D surface.

To represent a geographical entity, a coordinate system is needed. The most widely used reference system is World Geodetic System (WGS 84). It is a mathematical system used to plot locations globally. It is a standard coordinate system. This is the default system used by GPS devices to represent geodetic data. In the process of analyzing spatial data, the spatial reference system affects the calculation of physical features as different reference system provides different units of measurement.

Geospatial topology explores the rules concerning the relationships between the points, lines, and polygons that represent the features of a geographic region. Topology is also used for analyzing spatial relationships in many situations and the importance of topological relations is widely recognized [12] [13].

![Image of topological relationships]

**Figure 4: Examples of topological relationships**

There are various topological relations such as (Equals, Disjoint, Intersects, Touches, Contains, Within, Covers) that can exist between two geometries. The conceptual description of few of them is shown in the Figure 4. In this research we have applied these relationships to calculate the physical features of the world countries like: What are the neighbors of a particular country?

4. **GEOGRAPHIC DATA VERIFICATION**

This section demonstrates the research challenges that we have encountered during verification of data using maps. In order to verify the authenticity of the geographic data, the first step was to calculate the geographic features of selected countries from maps followed by their comparison with the geographic facts found in the CIA world fact book and Wikipedia. For this purpose we have downloaded the shape file (.shp) of the world countries from Natural Earth⁵ which is a great resource for geographic shape files. The SQL server spatial tool “shape to sql” was used to upload the shape files in a database. Afterwards spatial queries were applied to find out the topological relationships between countries on the map. The map of world countries is shown in the Figure 5.

![Map of world countries]

**Figure 5: World map showing countries**

```
1:Declare @g geometry
2:Select @g=geom from World_Countries where
3:country='Pakistan'
4:select country,geom from World_Countries
5:where geom.STIntersects(@g)=1
```

**Code Listing 1: Spatial query to find neighboring countries of Pakistan**

![Map showing neighbors of Pakistan]

**Figure 6: World Fact book Field Entry showing border countries of Pakistan**

![Map showing neighbors of Pakistan]

**Figure 7: Map showing neighbor countries of Pakistan**

![Map showing neighbors of Austria]

**Figure 8: Map showing neighbor countries of Austria**

⁵http://www.naturalearthdata.com/
Research Question 1: How to verify neighboring countries of a particular country?
To find out the neighbor countries we have used adjacency query. Figure 6 shows the field entry “Land boundaries” from CIA world fact book. The sub entry “border countries” reveals the fact that there are four countries that are sharing a border with Pakistan. Figure 7 verifies this information as the map indicates that there are 4 neighboring countries of Pakistan. This map is drawn using adjacency spatial query which extracts all objects in the map that are adjacent to a particular object in the map as shown in Code Listing 1. The map showing neighbor countries of Austria is also shown in Figure 8.

Research Question 2: How to verify boundary length of neighboring countries?
A boundary is a real or imaginary line that separates two things. The most obvious type of boundary is a physical boundary. A physical boundary is a naturally occurring barrier between two areas. A political boundary is a boundary line drawn between two countries. Political boundaries are created by humans, unlike natural boundaries which are created by natural processes leading to formation of rivers, watersheds, mountain ranges, or coastal lines. Natural boundaries have a pronounced obstacle impeding movement which may differ from one species to another. Humans usually ignore the natural boundaries when setting political boundaries. We are verifying the length of political boundaries in this particular geographic verification.

Research Question 3: How to verify area of a country?
The area of a country is one of the main geographic features. Russia is the largest country in the world with respect to area. To verify this feature we have taken the country Afghanistan as an example. The area of Afghanistan is found to be 652,000 sq. km approximately as shown in Figure 11.

Research Question 4: How to verify the boundary of a country?
A polygon is defined by a closed line string called exterior ring [1]. As countries in the map are represented by polygons, boundary calculation of countries becomes easier. The length of shared boundaries between Austria and Italy using STIntersects() function, as it returns the shared area that is common between two geographic objects which are countries in our case. The unit of length is calculated according to the reference system. To calculate the length in meters the SRID 4326 is used.

Code Listing 2: Spatial query to find the boundary length of Italy sharing border with Austria

1: select 8g=geom.MakeValid() from
2: World_Countries where name='Austria'
3: select country + ' ' +
4: cast(round(geography::STGeomFromText((geom.
5: STIntersection(@g).).STAsText(),4326)
6: .STLength()/1000,0) as varchar(10)) +
7: 'km' as
8: Label,geom.STIntersection(@g).from
9: World_Countries where
10: geom.STIntersects(@g)=1 and (coun
11: try='Italy')

Figure 11: World Fact book Field Entry showing area of Afghanistan
The map of Afghanistan shown in Figure 12 is showing the area in square kilometers, thus verifying the geographic entry “Area” in the CIA world fact book as shown in Figure 11.

Figure 12: Map showing total Area of Afghanistan
DECLARE @g geography
select @g = geography::STGeomFromText(geom.STAsText(4),4326) from World_Countries where Country='Afghanistan'
select 'Area '+cast(convert(varchar,cast(round((@g.STArea())/100000,0)as money),1) as varchar(10)) + ' sq km', @g

length of the outer boundary of the country is calculated using the spatial query listed in Code Listing 4. The STExteriorRing() method is used which returns outer closed Line String of the Polygon. The STLength () function is then used to calculate the length of the border of Belgium. The boundary of Belgium is found to be 1385 km thus verifying the entry in World factbook as shown in Figure 14.

Code Listing 4: Spatial query to find the boundary length of Belgium

Research Question 5: How to automatically verify basin countries through which a river crosses?
To answer this question let's take an example - the river Mur which is a big river flowing in Austria. We have drawn a map shown in the Figure 15 using Code Listing 5. This verifies the fact about the basin countries displayed in the infobox of Wikipedia article as shown in the Figure 16. It is clearly shown in the map of Figure 15 that river Mur, which is shown by a line string, originates in Austria and is flowing through Austria, Slovenia, Croatia and Hungary represented by the colored geometric elements in the map.

Code Listing 5: Spatial query to draw map to show countries through which river Mur crosses

One way of doing automatic verification is using STCrosses() method as shown in Code Listing 6. It returns true if a geometry instance crosses another geometry instance and returns false if it does not. This function is of extreme importance, as it finds out the countries through which a river crosses. One special care must be needed before applying this method, as it cannot be directly applied to geographic data types so first they are converted to geometric data types. For performing automatic verification of rivers crossing through countries we have taken two shape files one describing rivers and the other representing world countries. After converting them to geometric data type, we applied the STCrosses() method to verify the countries through which each river crosses, this leads to automatic verification of parameter the basin countries of rivers. In line 10 of Code in Listing 6 we can pass any river name and the query draws...
the map of basin countries for that particular river.

| 1:Insert into #temptable  |
| 2:select  |
| 3:name,geometry::STGeomFromText(geom.MakeValid()) from rivers  |
| 4:union all  |
| 5:select  |
| 6:NAME,geometry::STGeomFromText(geom.MakeValid()) from World_Countries  |
| 7:select @g=geom from #temptable where  |
| 8:name='Mur'  |
| 9:select @g=geom from #temptable where  |
| 10:STAsText(),4326) from World_Countries  |
| 11:select * from #temptable where  |
| 12:@g.STCrosses(geom)=1  |

Code Listing 6: Spatial query to automatically verify countries through which a river crosses

5. CONCLUSION

In this paper we explained the significance and use of spatial queries. We have applied a unique way of checking geographic data. Our efforts are mostly based on manipulation of world maps. This is a step towards automatic verification of data sources. We have examined the role of spatial operators and particularly linked their significance with the verification of geographic information with the real world examples.

ACKNOWLEDGEMENT

This research work was conducted as part of a Ph.D. thesis involving Austria-forum.org in the Institute for Information systems and Computer Media (IICM) of Graz University of Technology Graz (TU Graz) under supervision of Professor Hermann Maurer.

REFERENCES


AUTHORS

Rizwan Mehmood did his Masters in Information Technology (MS-IT) from NUST. He is currently PhD student at the Graz Technical University. He is the author of five papers and a member of the software development team of Austria Forum. Main research and project areas: Information architecture, multimedia information systems, Web based learning; languages; artificial intelligence; information retrieval; Web 2.0; semantic Web; database Geographic data mining.

E-mail: rmehmood@iicm.tu-graz.ac.at
How Open Content Servers Can Be Made Beneficial for Learning and Education

Ebner, Martin; Scerbakov, Nicolai; Schön Sandra

Abstract - The development of Open Educational Resources (OER) is necessary for the classroom of tomorrow. Especially in Central Europe the copyright law does not allow the use of not-licensed content located in the Internet for educational purposes. Therefore different projects and initiatives started to provide Internet platforms with free educational resources or links to such resources. For educators as well as learners searching for open content on several platforms can be very exhausting. In this publication we present a first prototype for mobile devices that allow users to find open educational resources in minutes. In the discussion, benefits and handicaps of the approach are pointed out. The research work shows that the application supports the daily life of teachers and learners.

Index Terms: Open educational resources, Open education, Mobile application, iOS, Android

1. INTRODUCTION

UNESCO announced its initiative “Free Educational Resources” in 2002. The UNESCO International Institute for Educational Planning (IIEP) defined the term “Open Educational Resources (OER)” as the „wish to develop together a universal educational resource available for the whole of humanity, to be referred to henceforth as Open Educational Resources“ [1]. The core of this issue is the main idea to make education and any content needed for educating people free available. One milestone for OER was the start of the “MIT OpenCourseWare” of the Massachusetts Institute of Technology (MIT) in 2002. Though MIT did it mainly for marketing purposes it is nowadays a very well-known and huge repository for lecture recording, videos, and additional content. Two further publications – a questionnaire about OER of the OECD in 2007 as well as a first draft about the OER-movement by the William and Flora Hewlett Foundation [2] helped to gather further attention. First OER projects which were co-financed by the European commission, started in the same year (OLCOS, Bazaar). Nowadays, there are many initiatives and projects dealing with OER; some of them are listed:

- MIT OpenCourseWare: Since 2002 the MIT offers free online courses and got a number of awards for its trend-setting initiative. In 2008 there have already been more than 1,900 courses from 33 different disciplines online [3].
- OpenLearn: The Open University UK also started a project, called OpenLearn founded by the William and Flora Hewlett Foundation. Due to the fact that in the meantime also Web 2.0 technologies became important they were integrated into the project following the idea that OER content can be enhanced by learner’s active participation [4].
- Wikieducator: At wikieducator.org a Media Wiki serves as one of the biggest OER-platfroms worldwide. The main idea is to provide resources about technology-enhanced learning and how to find and use Open Educational Resources.
- ZUM-Wiki: The ZUM-Wiki is the largest OER-wiki in German speaking countries aiming to offer open educational resources for schools. The repository for teachers is of mainly secondary level.

Of course this is only a short selection of many more different platforms in the World Wide Web [5]. With other words, the biggest advantage, which is to have no restrictions in offering OER, is also its worst disadvantage – it is hard to find appropriate material for teachers within a suitable short time span.

In this publication we deal with the way of how to put all these repositories together, so that teachers, lecturers, and educators can find appropriate OER content for their daily work. First of all the reason for using OER is pointed out, afterwards we describe a prototype in detail. In the discussion we focus on the constraints as well as the possibilities of our approach.

Manuscript received April, 2014.
Ebner Martin: Corresponding Author is with the Social Learning Department (as part of the Computer and Information Services), Graz University of Technology (e-mail: martin.ebner@tugraz.at).
Scherbakov Nicolai: Author is with the Institute of Information Systems and Computer Media, Graz University of Technology (e-mail: nsherbak@icm.edu)
Schön Sandra: Author is with Salzburg Research GmbH. (e-mail: sandra.schoen@salzburgresearch.at)
2. FROM COPYRIGHT TO OPEN EDUCATIONAL RESOURCES

In Central Europe a strict copyright law supports the rights of artists, musicians, and authors. Thus it seems of high importance to protect the rights of creative working people. This makes it nearly impossible to use such content for educational purposes. This is even the case when its educational usage is intended. But let us give a short example: If a university lecturer produces content for a lecture and provides it (online and digital) to the students to assist their learning processes, students are not allowed (!) to copy any of the content into their learning personal notations and share it with anyone per se. Each single student has to ask whether this is allowed or not. Of course, any other changes or even sending such content per e-mail to colleagues is simply not allowed. It can be summarized that learning content provided by an author can be only used in a kind of read-only way otherwise the copyright is being violated and the violator can be sued (which happens quite often). Reality is different: we have in that sense every day lots of incorrect copyright behavior that misleads teachers and students to hide their content from each other.

Open educational resources are exactly the opposite of this development and therefore a solution to overcome these problems. They are not only freely available, but also free to use. Each single resource is delivered with a license that allows the usage by teachers as well as by learners in a defined way. "Open" means that [6]

- It is available for free,
- It is useable for free (can be changed, remixed, ...),
- It is possible to use and modify the material with free available software (e.g. OpenOffice), and
- It supports open teaching and learning processes.

Beside the copyright law Geser [7] pointed out further several benefits of using Open Educational Resources in education (p. 21):

- OER offer a broader range of subjects and topics to choose from and allow for more flexibility in choosing material for teaching and learning.
- OER leverages the educational value of resources through providing teachers personal feedback, lessons learned and suggestions for improvements.
- OER provide learning communities, such as groups of teachers and learners, with easy-to-use tools to set up collaborative learning environments
- OER promote user-centered approaches in education and lifelong learning. Users are not only consumer of educational content, but create own materials, develop e-portfolios and share study results and experiences with peers.

Since those early days of the OER movement, different publications pointed out why OER is highly relevant also for Higher Education [8] [9] [10] [11]. For example, the necessity of an own OER strategy is carried out by Schaffert [12] and executed for the first time at Graz University of Technology [13].

3. PROTOTYPE

As described before, Open Educational Resources are a prerequisite for teachers and learners of tomorrow and it is still hard to find appropriate resources for special teaching or learning behaviors. Most resources are parked on Wikis using the software Media Wiki, without any national or international strategy, as well as using no Meta standards. Therefore it is nearly impossible – Wikipedia seems to be the only exception – to find these platforms on the World Wide Web or their content without any knowledge about their existence.

3.1 Creative Commons License

In the United States of America it is possible to provide any content as so-called “Public Domain”, which can be used by anyone for anything. The copyright law in Central Europe did now allow defining any content as public domain. The author of any content owns the copyright for at least 70 years after his/her death per law. Therefore it is absolutely necessary that authors define how their content can be used. In general, there are numerous different license models, but especially in the field of OER the so-called Creative Commons licenses (CC) are common and widely used. The CC model benefits from the fact that there is a standardized, easy understandable description about how content can be used. Such content is signed with the “CC”-symbol; additionally to the "CC"-symbol the following symbols and options can be added and combined:

- BY: In the case of re-use of the content, the author has to be named.
- ND (“no derivation”). It is not allowed to change the provided content in any case.
- NC (“not commercial”): It is not allowed to use the content for commercial purposes (for example if students have to pay for a single course).
- SA (“share alike”): In case there is any change to the content it must be republished under the same conditions, which means the same license.
For example, the world best-known encyclopedia Wikipedia is using the CC license CC-BY-SA-3.0 (as shown in Fig. 1).

Fig 1 Creative Common License CC-BY-SA

3.2 Technical solution

As described, it must be taken into account that there are a number of servers holding different Wiki-systems for Open Educational Resources. Due to the fact that more or less all those servers are using today’s most favored system Media Wiki, the idea came up to send one search request for a particular search query to all the servers in parallel. With other words it should be possible to search in different systems in real time. Furthermore, such a search is often done when travelling and therefore an application for mobile phones should be developed. The prototypes are realized as two mobile applications for the two most spread mobile operating systems – iOS and Android.

Fig. 2 Main Screen

Due to the requirement that one search query is sent to different Media Wiki instances the offered Application Programming Interface (API) of those Wikis is used. According to the official documentation of the API [14] it is possible to send an http-request to the Media Wiki that allows searching for data, extract data, change data, login or logout, and even more on the corresponding system. In our particular case just three requests are necessary:

- Open Search: The corresponding Wiki system gets a search request and gives back an answer string with a maximum of 10 related articles
- Parse: If the user decides to read one of the related articles in detail, the app asks for the whole article.
- Random: This is a fun mode of the application – when users click on random functionality they get randomized articles from the platform.

3.3 The Mobile Application - How it works

The prototype has been developed for two different mobile operating systems:

- iOS: One application was programmed using Objective C and Apple’s developer environment Xcode. The final app is running on iPhones as well as on iPads.
- Android: The second application was programmed using Java and Android’s Software Development Kit (SDK). The final app is running on all Android-based mobile devices with Android 2.3 or higher on board.

Before programming was started so-called paper mock-ups were carried out aiming to guarantee a sufficient user experience and that both applications are using the same program logic as well as look and feel. The following description of the main functionality of the applications is the same for both apps; the screenshots are taken from the iOS version.

Fig. 2 Main Screen

Fig. 3 Search results

Fig. 4 Article

Fig. 5 License Screen

Fig. 2 shows the start screen – it’s simply a search field with the intent that the users provide a keyword they are looking for. After tapping on the search button, numerous search queries are sent to the wiki APIs defined in the preferences (there are 12 different Wiki systems predefined). Fig. 3 lists the search results in a table view, which article has been found in which Wiki system. In our example there are different results in the Wikipedia (English and German version), the Wiki Educator and in the Creative Commons system. If any article of the list of results is chosen, Fig. 4 is appearing. This screen shows the detailed article as well as the according license at the bottom of the page. In our particular case the license is a creative common license,
CC-BY-SA. If users taps on the license icon they will get the appropriate licenses description (if available) as shown in Fig. 5. All copyright issues are listed and point out the way this content can be used in education.

Finally Fig. 6 shows the left preference side bar, which can be accessed by a swipe gesture. There the user can choose the random modus, save bookmarks, a general description about the CC licenses, a short “about” window, and the possibility to share the article via e-mail or social media. The last item “Wiki” gives the users the possibility to add numerous Media Wiki systems on their own or to delete existing ones.

The screenshots for the Android version are exactly the same.

4. Discussion

In this section we discuss important issues of the prototype. After launching the apps in the appropriate online stores it was tested with several users of different backgrounds – educators as well as students. The following outcomes are worth to be mentioned:

- The apps are an easy and fast way to search for OER-content. The users had no problems in using the app and found very fast appropriate OER content.
- According to some users, the app will support their daily life, due to the fact that they can search for the same keyword in many different wiki systems in parallel. Furthermore, they suggested us to predefine some more systems.
- Some users mentioned that the app is not for beginners, due to the fact that the licenses are shown as symbols. If a user never saw these symbols before he/she will not be able to understand it intuitively.
- The app can be extended to any other Media Wiki system just by providing the URL. Other platforms are not supported.
- Some users complained about that there is no desktop version for Windows or other operating systems.
- One problem of the applications is that not every found article had a license. From a technical point of view this occurs when the Wiki system in general provides no licenses or when the license is not defined in the meta data of the article.

In this research project we described a first prototype that should help educators as well as learners finding content for their daily work with an open license. Bearing in mind the evaluation process, we were able to bring different open content servers together.

5. Conclusion

Due to the fact that during the evaluation process some teachers complained about the necessary pre-knowledge about licenses in general, in the meantime a further application has been programmed. That second app (currently only available for iOS-devices) called OER-Remix Game gives lecturers as well as learners the ability to exercise their knowledge about OER and licensing. The app itself is defined as a game to encourage learning in a playful manner.

Further research work has to address the integration of further platforms beside Media Wikis. From a technical point of view this is not a trivial approach, due to the fact that some platforms did not offer any API or a complete different one. Nevertheless, this is one of the core future tasks in assisting world educators with appropriate open content.

Finally it can be pointed out that we have to bear in mind two main tasks for servers providing open content. First of all, each single content should be offered together with a license, defined in the meta data. Second each open content server should provide an API which allow to search and find OER.

Acknowledgment

We would like to thank Alexander Grabner and Daniel Karner who programmed the prototypes for iOS and Android and their nearly endless patience during the evaluation process.

References


A Review of the Open Educational Resources (OER) Movement_BlogLink.pdf (last access: April 2014).


AUTHORS

Martin Ebner is currently head of the department Social Learning as well as senior researcher at the Institute of Information Systems and Computer Media at Graz University of Technology. For further details visit his homepage http://www.martinebner.at

Nikolai Scerbakov is a professor at the Institute of Information Systems and Computer Media at Graz University of Technology. More info is available as http://coronet.iicm.edu/Dbase1/nick1.htm.

Sandra Schön is senior researcher at Salzburg Research Forschungsgesellschaft (Austria) within the Innovation Lab. Her focus lies on innovative technologies for learning. More in her Weblog: http://sandra-schoen.de
Digital Archives as Part of Digital Humanities Research Infrastructure
Towards a Standardized Model of Archiving and Dissemination

Stigler, Hubert Johannes

Abstract - Due to the increasing degree of digitization in research, the topic of modeling scholarly content is gaining increasing prominence in Humanities and Cultural Studies. XML-based data formats appear well-suited to flexible, metadata-enriched forms of storage of textual data: the primary content of documents is augmented with additional descriptive elements, based on modeling standards like TEI (Text Encoding Initiative), OWL (Web Ontology Language), or SKOS (Simple Knowledge Organization System). These standardizations lay the foundation for the semantization and, consequently, the automated processing and analysis of specialist knowledge, incorporating domain-specific ontologies and vocabularies. The separation of content and its presentation as a fundamental feature of XML-based (eXtensible Markup Language) formats implies a high degree of flexibility when dealing with the analysis and transformation of the original (textual) data in different presentation forms, but also calls for standardized workflows in the processing of such data. A variety of solutions and frameworks have been developed for this task. This contribution aims to present and discuss an object-oriented approach to a digital archive, which was implemented in a FEDORA (Flexible Extensible Digital Object Repository Architecture) -based digital repository environment.

Index Terms: Information modeling, Long-term preservation in the humanities, Text Encoding Initiative

1. INTRODUCTION

Issues of management and delivery of digital resources are nowadays gaining in importance in museums, archives and libraries, but also in humanities research. In many places, the digitization of collections has begun as an attempt to preserve and considerately make available books, manuscripts and other cultural artifacts. The currently available storage media hardly sets physical limits to mass digitization projects, but there are issues of logistics and sustainability that arise in such projects in many ways.

Thus, as early as 2011, the German Science Council spoke out in favor of the expansion of sustainable, research-suitable digitization and efforts to coordinate the standardization and interconnection of relevant IT-infrastructures.

2. SEMANTIC ENRICHMENT AS A KEY FOR A NEW METHODOLOGY

Such research-suitable digitization reaches far beyond a purely pictorial and textual representation of cultural artifacts in the computer. It is not only about digitizing source materials and other scientific resources in the classical sense (colloquially: scanning them), i.e. converting them from an analog form of representation into a – in relation to the data format usually proprietary – digital form. Sustainability thereby only arises through the enriched representation of sources, traditions, texts and images aimed at the formal visualization (explication and contextualization) of the semantic structures contained in such data; hence, the enrichment of the content of the digital objects emerging during the process of digitization with (standardized) metadata, referring to different levels of description, such as logical text structure, interpretative or narrative levels, morphology, syntax, etc. Through such enrichment, collections of digitized objects evolve into representations of digital editions: formally enriched cultural artifacts become the empirical data base for research in the humanities by opening up new possibilities for an IT-based representation and analysis of these semantic structures: for example, paleographic properties of a manuscript can be formally differentiated and statistically analyzed, narrative levels or regional references in a literary text corpus can be visualized and opened to interpretation, extensive image collections can be subjected to comparative analyses [9]. The main difference to a paper-based edition is the effort to enrich the edited text with metadata which does not only refer to descriptive (i.e. describing an electronic resource) data about data, but also to
information which semantically structures and analyses the content of a text document.

Plain-text of a document produced with a word processing application or a transformed PDF file of such a document is from this perspective merely an amorphous mass, whose structuring refers to the lowest unit of characters (spaces, punctuation, etc.). Simple forms of automated enrichment with semantic meaning can be applied at this stage – for example for the recognition of sentence and word boundaries – as well as algorithms which automatically annotate the morphosyntactic properties of words for later use in search processes whose query language allows the search for grammatical structures in a text database.

So-called repositories or digital archives constitute the necessary IT-infrastructure to keep the data of digital editions sustainable and permanently available. They provide a quotable provision of the available content and thereby organize the persistent survival of information in an ever-changing technological environment. Many national and EU-wide projects and initiatives (CLARIN, Common Language Resources and Technology Infrastructure, http://www.clarin.eu; DARIAH, Digital Research Infrastructure for the Arts and Humanities, http://www.dariah.eu among others) are currently evaluating and developing such systems which are – generally speaking – based on the intention to improve scientific research and communication processes with uniform access to electronic knowledge pools and make them more transparent (to outsiders).

3. DIGITAL ARCHIVES AS VIRTUAL RESEARCH ENVIRONMENTS

Digital archives are not merely storage spaces, but also support the processing of digital resources in different scientific scenarios. Flexible authorization models control the web-based access to source materials and research results. Long since, the design of relevant applications is no longer exclusively oriented on the concept of a pure data-store, i.e. an optimized storage and retrieval facility for static content (text, picture, and sound or film documents). In many projects, modeling standards are being defined and workflow models implemented which aim at the digital representation of the entire creation process of scientific research results (e.g. the eSciDoc project at the FIZ-Karlsruhe, https://www.escidoc.org, or Text Grid, http://www.textgrid.de/). Issues of text-, or better, information-immanent annotation and thus the semantization of the content of texts, images, movie clips, etc. come to the fore. Digital data, manually or (semi-)automatically enriched with domain-specific metadata (e.g. text corpora containing lemmata or morphosyntactic information, structurally annotated transcriptions of manuscript, etc.) can not only be researched in an intelligent and ontology-based way but also constitutes important points of reference for empirical analyzes and thus supports the theory construction of the respective scientific domain. This development also renders the classic division of labor between producers and archivists of scientific research results obsolete. Issues of digital archiving or more generally the digital representation and modeling of knowledge are becoming relevant methodological issues in the scientific domain of the respective contents. Hence, in relevant application scenarios, workflows which allow for the collaborative editing and managing of digital resources are to be preferred.

4. THE ‘LINGUA FRANCA’ FOR DIGITAL ARCHIVES

If the enterprise digital preservation of scientific and cultural heritage is to prove a sustainable success, it is necessary to establish new avenues for long-term preservation scenarios beyond proprietary data formats: These are too short-lived for sustainable forms of archiving, and some may require error-prone, automated migration processes in the digital archive. Overall, XML-based data formats have prevailed in recent years, not only as a format for descriptive metadata, but also for the overall modeling and annotation of the content of digital objects. As a text format in the humanities, the TIE (Text Encoding Initiative) metadata set offers very flexible and comprehensive options for the human-readable modeling of text documents of almost any origin, based on the premise of separation of content and representation [14]. Overall, it can be observed that the progressive dissemination of XML-based data formats has resulted in the increasing development of XML technologies and tools in applied IT-areas which are also useful for knowledge modeling.

An essential prerequisite for the establishment of XML – far beyond the original application field of (syntactic) information structuring – was the stabilization and consolidation of the standards, especially by (a) the normalization of the structure of well-formed XML documents through the introduction of the XML Information Set standards, (b) the introduction of an extensible type system in XML, which allows the description of arbitrary data structures (XML Schema) and (c) the establishment of a uniform convention for the use of namespaces [4,5]. Thus, the foundation was laid for the development of
complex annotation languages, as represented by the TEI in its present version P5. This version was developed with the intention to construct a universal convention for text annotation and record it in the form of guidelines (applicable to all languages and text types) [3]. These guidelines provide a flexible framework for the definition of (normative) encoding standards whose application areas include sources and documents as they exist or are being produced in a variety of (humanities) disciplines, from historical documents to texts generated in survey situations of quantitative and qualitative empirical social research, to literary and linguistic text corpora [11,12]. In addition to continuous texts, TEI can also be used for the annotation of non-continuous texts such as dictionaries, etc.

Complementary to this primary application of XML-based information modeling, a number of other functional areas in information processing have emerged in which XML plays a central role in digital preservation:

(a) Metadata description and knowledge management, i.e. the secondary use of XML to add additional description elements to content, e.g. by using standards like the Resource Description Framework (RDF, http://www.w3.org/RDF/) [2], Topic Maps (http://www.topicmaps.org/) and Web Ontology Language (OWL, http://www.w3.org/TR/owl-semantics/). These standardizations provide the foundation for the modeling of domain-specific vocabularies as well as making them accessible to automated processing (e.g. in a search process). Extensive examples of such domain-specific ontologies can be found in the field of archival and museum work: the International Committee for Documentation of the International Association of Museums (CIDOC, http://cidoc.ics.forth.gr/) has compiled an extensible ontology for terms and information in the area of cultural heritage in the CIDOC Conceptual Reference Model. The SKOS specification (Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/) presents a first-time effort by the W3C to standardize a formal language for encoding documentation-languages such as thesauri, vocabularies and other controlled vocabularies based on the Resource Description Framework.

(b) Transformation of information, i.e. the use of XML standards to map information structures to each other, e.g. to derive a representation format from an XML structure. Here, Extensible Stylesheet Language (XSL, http://www.w3.org/TR/xslt20/) affords special consideration as the umbrella term for a complex control system, which consists of three specifications: (1) XSLT (XSL Transformation), a transformation language for the structural editing of XML documents which describe a rule-based transformation process from an input file into one or more output files of any target format using XML syntax, (2) XPATH (XML Path Language), which allows the selection of (virtual) sub-trees of an XML tree structure, and (3) XSL-FO (XSL Formatting Objects), a standard for printed page description.

(c) Exchange of information, which is the use of XML as a universal data exchange format between applications, also at the level of protocols over the internet.

(d) Application modeling language, that is the use of XML in the design, programming and deployment of applications, e.g. in UML-based (Unified Modeling Language) development environments, but also as a control-relevant modeling language for process sequences in web-based application frameworks.

5. An XML-based Publication Framework

For the processing of XML-based data, a number of specific publication frameworks exist: They support, among other things, the representation of different aspects of a multi-layered, modeled text structure in different formats (or views) for scientific end-users. In general, they are characterized by the following features: (a) realization as a client-server application based on W3C standards; (b) use of a Three-Tier-structure, where the client is usually a web browser and a corresponding server-sided application logic is implemented; (c) modular integration of XML-processing components (XSLT processor) and (d) separation of the (XML) content from its presentation aspects to achieve flexibility with respect to the output generation and processing of the data.

As an open source project that implements this requirement profile, the COCOON framework (http://cocoon.apache.org) developed by the Apache Software Foundation has gained wide acceptance. COCOON is a Java-based, dynamic component framework that can be integrated into a JSP (Java Server Pages) container environment and implements certain standard components alongside its own modules for controlling XML-processing. The open, component-based architecture allows for easy integration of database- or authorization-modules. Like almost any web application, COCOON is embedded in a fixed request-response-cycle. A request in COCOON is implemented as a pipeline of successive steps. Within COCOON, the URI of a request is
analyzed and passed on for processing to one of the pipelines defined in the central control file through a so-called Matcher. Each step of the procedure can write data depending on the current pipeline state and read from various data sources (file, [XML-]database, web service, etc.). With this simple approach, complex tasks can be disassembled into several subtasks. Communication within a pipeline occurs via so-called SAX (Simple API of XML) streams, in which responses of a pipeline component are passed on to the next component in the processing chain in XML format. In the last step of the pipeline, the output which is transmitted back to the client is generated (where again, XML components can take over the serialization of the XML stream). Standard components of the framework are thus well suited for the realization of (multilingual) web applications with underlying XML-based data and text bases.

6. ASSET MANAGEMENT SYSTEMS

So-called asset management systems offer a framework for digital archives going far beyond the functionality of COCOON [1]. The term asset in this context describes the smallest unit which is structuring, described, and managed by the system, comparable to a catalog entry. Such an asset is made up of a primary data stream (e.g. text document, spreadsheet, presentation file, audio or video file etc.) and at least one set of descriptive metadata (Dublin Core). Such so-called simple model assets stand opposed to compound model assets which can consist of a wide range of primary data streams and associated functions: e.g. an asset for digital books, consisting of all image files of photographs of a manuscript, the edited text (in any text format) and a set of methods which allow scrolling through the pages of this (virtual) book or zooming details of the individual pages. They are used quite generally for the storage and management of digital resources, as they occur in scientific contexts in great variety. In contrast to content management, asset management puts special emphasis on aspects of sustainable, metadata-based and quotable archiving of and access to digital resources, controlled through flexible permission models. Sustainability in this context refers to the long-term availability of resources, but also aims at fundamental considerations in connection with archiving projects, such as recommended (and actual) data formats. In the context of scientific data management, it is paramount to ensure that – regardless of changing software environments – texts, pictures, movies, statistical data bases and other materials archived for scientific purposes remain quotable and safely accessible over longer periods of time (comparable to print publications) [15].

As basic functions to accomplish its tasks, an asset management system provides import and export operations for data sources (possibly connected with format conversions, such as from MS Office to XML), the option of enrichment with different descriptive but also administrative metadata, the pooling of resources in containers and the versioning of data sources, as well as strategies for the URL-based addressing of individual sub-entities of an asset.

7. FLEXIBLE EXTENSIBLE DIGITAL OBJECT DEPOSITORY ARCHITECTURE

Furthermore, an asset management system as described here is geared towards realizing the paradigm of Single Source Publishing: content can occur in a variety of representation formats which are dynamically generated (when referencing the content through a web-browser) from the content stored in an asset using so-called style sheets. When searching for a suitable platform for the realization of a suitable IT-infrastructure for the long-term preservation and archiving of digital (research) data at the Center for Information Modeling – Austrian Centre for Digital Humanities, extensive research ultimately led us to the open source project Flexible Extensible Digital Object Repository Architecture at Cornell University (FEDORA, http://www.fedora-commons.org).

Our own project with the acronym GAMS (Geisteswissenschaftliches Asset Management System, http://gams.uni-graz.at) is a digital archive for the metadata-based management and sustainable provision of digital resources. From manuscript to video, from text edition to image archive, it offers to faculty and students the means to archive and publish such resources in a standardized, quotable and web-based way. Heterogeneous requirement profiles are indicative of application scenarios of digital archives in scientific contexts: the specific areas of operation of the system range from learning object collections to digital editions, from video and film archives to (morphosyntactically) annotated and multimodal corpora. GAMS is also integrated in several European research infrastructure projects and is therefore not a mere island solution. As an example, the content managed through GAMS is integrated in Europeana (http://europeana.eu), a long-term EU project with the aim of initiating and implementing a common European search portal for scientific content.

On a structural level, the process of digitization also requires further thinking on sustainability. The European research initiative DARIAH aims
at the establishment of a sustainable digital research infrastructure in the university environment throughout Europe. Based on such research infrastructures, the sharing of resources, methods, data and experiences will be encouraged and scientists will be supported in establishing collaborative and digital research cultures to respond to their genuine research questions in new ways and even develop new research questions. Strategically, these projects aim at joint software development as well as the construction of competence centers to create the necessary institutional prerequisite for the further digitization of humanities.

At its core, FEDORA provides a database-driven, modular expandable storage and management structure (repository) for arbitrary (distributed) digital resources, with web-based access, guided by the principles of a Service Oriented Architecture (SOA) with the following properties [10]:

(a) Web-based (SOAP, Simple Object Access Protocol, http://www.w3.org/TR/soap/), platform-independent, distributed system architecture,

(b) Apache Lucene-based full-text index and versioning management of the asset contents, (http://lucene.apache.org/)

(c) RDF-based triplestore with the SQL-like query languages ITQL (Mulgara Project, http://docs.mulgara.org/tutorial/itql.html) and SPARQL (SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/).

(d) Definition of intricately controllable access rights to assets and their sub-entities based on XACML (Extensible Access Control Markup Language, http://www.oasis-open.org)

(e) Standards-based import and export formats: METS (Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/) and others,

(f) Unique URL-based addressing of digital resources,

(g) Support of standardized protocols for metadata exchange, like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) and others (http://www.openarchives.org/),

(h) The option to realize system environments with corresponding numbers of concurrent users by repository clustering and load balancing thanks to the carrier technology (Apache Tomcat, http://tomcat.apache.org/) [9].

As one of the most important paradigms in the field of software design, Wolf – referring to Gamma and Pree [16, 13] – identifies the formation and implementation of the concept of object orientation, which, in concert with the introduction of modular software concepts with high granularity and the systematization and standardization of object- and component-oriented software development through design patterns, has led to a high degree of reusability of software technologies.

Object orientation is characterized, among others, by the following properties:

(a) The definition of classes with associated properties (attributes) and methods as essential structural elements and

(b) The formation of class hierarchies through the principle of inheritance, taking advantage of polymorphism.

These principles are implemented in FEDORA not only at the level of system development, they also structure the application logic at the user level: Through the design of content models (object classes), complex object class hierarchies can be constructed in a FEDORA-based data repository [7,8]. Content models describe not only the content structure for an asset class (data streams) and the potential relations to other objects (container assets), but can tie so-called disseminators (methods) to the data of an asset through the use of Web Service Description Language (WSDL, http://www.w3.org/TR/wsdl.html): e.g. XSLT transformations which convert the XML data streams of an asset to any desired target formats (HTML, PDF, etc.); methods that convert a color image stored in an asset, into a black-and-white variant for the use of the image in offset printing; functionalities allowing the navigation in an indexed video file, and other features. With respect to an object model specifically geared to modeled text corpora, which is initialized at its instantiation with an XML file encoded in TEI format, these could be dissemination methods which in one instance represent an XML data stream as a navigable HTML document (e.g. with specific interactive analysis options, like the configurable highlighting of certain text structure levels), and in another instance as a PDF or LaTeX file.

FEDORA also supports the assignment of intricately configurable access rights to assets and their dissemination methods. Basically, all entities which are part of an asset can be individually addressed and referenced (web-based). Through XACML, individual access channels can be linked to distinct authentication and authorization rules. For example, all access paths to a text object except that which produces the representation in HTML format can be restricted to authorized project staff. FEDORA also supports standards such as LDAP (Lightweight Directory Access Protocol, http://www.openldap.org), Shibboleth (http://shibboleth.internet2.edu/).

Through this feature, related object data can
be stored in a common administrative and storage context (asset) while at the same time implementing query and editing scenarios controlled through differentiated access models.

8. CONCLUSION

Standardized annotation languages and technologies for the processing of XML-based data structures based on relevant reference models nowadays form the basis for the realization of sustainable long-term preservation and archiving scenarios for research data in the humanities. In such contexts it is a requirement to provide distributed digital resources through centralized storage, management and retrieval structures and thus ensure the quotable archiving of digital knowledge bases on the premise of recyclability. Although much has already been accomplished to that end, most of the existing technical solutions are essentially prototypes with a low degree of standardization. Many technical issues have been resolved, yet several desiderata remain:

(a) The development of standardized modeling languages for the description of processing workflows and forms of representation,
(b) An open methodological discourse on issues of the digital transmedialization of research data in the humanities,
(c) Adequate and standardized tools and corresponding technical infrastructures
(d) Trusted digital archives and sustainable institutional infrastructures.

REFERENCES

Thematic Digital Libraries vs. Wikipedia's "One Size Fits All" - Lessons Learned

Barthel, Simon; Balke, Wolf-Tilo

Abstract - Today the Web serves as the central information hub for almost all areas of daily life, where issuing a simple keyword query offers access to a dazzling array of information. And this does not only cover trivial information: also in the area of professional or scientific information there is a clear trend towards easily accessible information provisioning via the Web, e.g. in digital libraries, open access journals, or topically focused forums or newsgroups. But here, querying is mostly much more complex and offers a vast variety of community-specific interfaces, specific indexing schemes, and metadata-based access. This increased complexity leads to the question whether such effort is really needed or if general purpose knowledge portals, like for instance Wikipedia, would already be sufficient even for sophisticated tasks with a clear thematic focus. In this paper we explore the challenges and chances of specialized thematic digital libraries reviewing typical use cases from different disciplines like chemistry or mathematics and argue that although one size does not fit all, there is a lot to learn from general purpose portals.

Index Terms: Digital libraries, Metadata generation, Metadata indexing, Wikipedia

1. INTRODUCTION

LIBRARIES have been in existence for over 5000 years: from the oldest known collections of cuneiform written on clay tables in Sumer about 3000BC via the great library of Alexandria, where for the first time the idea of collecting all scholarly knowledge of the world in a single place was implemented, to today's central national libraries like the US Library of Congress containing more than 32 million books, 61 million manuscripts as well as millions of newspapers, scholarly articles, microfilm reels, comic books, and maps in more than 400 languages. With growing collections it became obvious that for supporting advanced search tasks and for managing the heterogeneity in this amount of data sophisticated schemes for bibliographic indexing and topical annotations were needed. Prime examples are the Dewey Decimal Classification (DDC) or the Library of Congress Classification (LCC). But already at an early stage there was a distinction between generally applicable universal schemes like DDC and LCC and more expressive specific classification schemes for particular subjects or types of materials like for instance the National Library of Medicine classification system (NLM).

With the increasing trend of digitization for improved accessibility the basic ways of describing library items slowly began to change. On one hand increasingly powerful computer systems enabled a self-description of documents even in large corpora (e.g., full text search, information retrieval, and text mining) as opposed to expensive manual indexing. On the other hand after the advent of the World Wide Web the active interlinking of documents and their subject-specific compilation in focused Web-accessible collections enabled powerful application scenarios like shared research environments or workplaces.

As far as information search is concerned both kinds of basic indexing schemes left their mark in digital collections and knowledge portals. Search engine technology of the universal type uses simple keyword search with only a modicum of general classification information like e.g. in Wikipedia. Navigational or facetted interfaces as often used in thematic digital libraries use specific classifications like subject-heading systems, taxonomies, thesauri, or any other type of structured controlled vocabulary.

But in contrast to the obvious advantages of higher expressiveness in querying using specific classification schemes in physical collections, where search was necessarily only based on metadata, the advantage of specific classification schemes in digital collections, where text-based indexing is easy and classification based on machine learning is possible, is an active area of discussion. Especially the growing heterogeneity of user communities due to increased interdisciplinarity and the rising costs for the classification schemes’ maintenance may be good points in favor of abandoning subject specific classifications leading to less complex information access. The exponentially growing amount of information to be searched on the

Manuscript received April 14, 2014.
S. Barthel and W.-T. Balke (contact author) are with the Institute for Information Systems (IfIS), University of Braunschweig, Germany.
(e-mail: {barthel,balke}@ifis.cs.tu-bs.de).
other hand may call more than ever for subject-specific methods for filtering and retrieval, where ease of access is sacrificed for search effectiveness.

In this paper we present lessons learned from designing search interfaces for thematic digital libraries in different fields like chemistry, mathematics, and medicine.

2. Challenges for Today’s Digital Libraries

Since the 70s lots of research has been done in the area of Information Retrieval and Machine Learning and the continuous improvement of processing power and networking capacities smoothed the way for big search engines like Google, Bing, Yahoo, etc. to become the first choice for casual users who wants to satisfy their information need.

Still, for scholarly users there are some specialized products like Microsoft Academic Search, Google Scholar or CiteSeerX, which can be used for literature search in arbitrary domains with acceptable quality of results. In contrast to general search engines here the document base is more restricted and some advanced features like author search or document ranking by numbers of citations are supported. In fact, today these products are the biggest competitors for digital libraries. Indeed it seems most natural to use e.g. Google Scholar for scientific searches, if one uses Google Web Search for everyday queries. If the results are satisfying, there is no reason why researchers should bother to constantly switch between several focused digital libraries and manually integrate information for a hypothetical improvement of search results. So the question is what exactly is the benefit of a digital library?

A number of international initiatives are already building vast digital archives of global content for universal access. For encyclopedic knowledge Wikipedia is the major resource worldwide offered in a vast variety of languages. For cultural heritage the two largest initiatives today are the Open Content Alliance and the Europeana, both of which already provide collections of several million curated and quite well-maintained items. However, today both are basically huge databases storing digital items and allowing the retrieval based on metadata. In particular, these systems only offer unitary navigational access to different disperse segments of information, although providing links to other related content. This makes the process of building a holistic view and a deep understanding of some topic (or information need) difficult and time consuming.

In contrast digital libraries spend enormous efforts on manually curating their collection and enabling customer-centered access. Every item is carefully indexed and enriched with metadata - e.g. structural elements like subject headings are annotated, authors or publishers are identified, semantic keywords are assigned and much more. The results of a query can then for example be filtered by the annotated metadata using a facetted search. Another aspect that distinguishes general purpose search engines from digital libraries is the quality management regarding the indexed items. For digital libraries it is important that every item corresponds to a real publication produced by a real publisher or by a real author while a search engine just crawls and indexes everything of potential relevance that appears to be a scientific paper or a book.

This quality management becomes more and more difficult. Consider for instance TIB Hannover, the German National Library of Science and Technology: currently about 90,000 metadata entries have to be manually annotated each day with strongly increasing tendency. Moreover, there are not only classical publications to be indexed, but also a growing amount of primary research data in heterogeneous formats ranging from experimental data sets via simulation data to descriptive models. However, as libraries have to annotate more and more data with the same amount of financial resources, the problem of retaining the high quality of annotations becomes more pressing from year to year. One way to cope with this problem is to focus closely on one domain and build dedicated subject-oriented digital library where the range of metadata values is limited.

The benefit of such a thematically narrow library is not only that due to the special focus less data items have to be catered for, but also that it can be more responsive to the special requirements of the domain in the sense of value-added services. For example depending on the user's expertise information can be provided in different abstraction levels or search interfaces can be tailored with respect to the domain.

The problem that the information need of arbitrary users cannot be satisfied by one source can be well illustrated with an article about the Higgs Boson published by Scientific American [1], which caters to quite a heterogeneous readership. The magazine tried to answer the question of the nature of the Higgs Boson in three different ways:

- With an introductory text understandable for the general reader featuring 2 paragraphs length provided by an author from the Northeastern University
- With a more fundamental text of 6 paragraphs provided by an author from the Santa Cruz Institute for Particle Physics
With a text aiming at experts in the respective domain over 9 paragraphs provided by an author of the Fermi National Acceleration Laboratory.

Clearly, performing such an amount of journalistic editing to satisfy the information need of every possible target user group for all topics cannot be the default case. A stricter focus on the other hand combined with domain specific personalization techniques as provided by most subject-centered digital libraries has great potential to provide users with the exact level of information that matches their information need and expertise.

Also, when performing research in special domains there are lots of search requirements and access patterns that only apply to that domain like e.g. querying models, experimental raw data, test corpora or structured domain knowledge. But whenever a search interface for a domain independent library is provided, there are two extreme choices: either all kinds of access can be supported whether they are applicable to all items or not, or the interface can only offer access using common entities that are available in every scientific domains, like e.g. authors, articles, or books. In the first case the interface becomes cluttered up to the stage of being unusable, in the second case search expressiveness is severely hampered. Thus, most interfaces aim at a more or less satisfying compromise.

In the domain of chemistry researchers might for instance be interested in an interface that understands a chemical formula, so that they can perform a query like C_6H_6. They will then certainly find out that C_6H_6 corresponds to the molecule benzene – and 217 other molecules that all have 6 hydrogen and 6 carbon atoms but are structured in different ways. The important question that has then to be answered by the retrieval system is: what molecule is most relevant to the researcher? Since molecules are not per se relevant or irrelevant, this question is impossible to answer until the chemist’s research context or search context is known. Also, this search context and the entities in the database have to be represented semantically to determine how strong an entity matches with a search context. This is normally done by using semantic metadata.

Semantic metadata, in the form of structured domain knowledge (like e.g. ontologies, taxonomies or controlled vocabularies) exists for all kinds of scientific domains, like the Mathematics Subject Classification (MSC) for mathematics, the Open Biomedical Ontologies (OBO) for biology or the Medical Subject Headings (MeSH) for medicine. Once the entities of interest are annotated with the respective domain knowledge with high quality and a search context is also represented using the same form of structured domain knowledge, it is possible to rank entities with respect to the user’s context.

In the chemical domain the task of annotating such semantic metadata is e.g. performed by the Chemical Abstracts Service (CAS) with a huge amount of effort. The database containing these annotations is however not freely accessible but a standard CAS user license costs around 30,000 USD. For researchers, the purchase of such a user license can of course be a burden. Fortunately, not only commercial services provide semantic annotations but also digital libraries like e.g. the Zentralblatt MATH (zbMATH) which is the biggest digital library for mathematics in Europe. Every article being published in the zbMATH is annotated with respect to the MSC that is also maintained by the zbMATH in cooperation with Mathematical Reviews. However, the increasing amount of digitally available data becomes a continuously growing problem for digital libraries as more data needs to be annotated with the same amount of financial resources.

A solution to this problem might be to use automated text categorization approaches to automate the process of indexing. An extensive large scale study of state-of-the-art machine learning technologies applied on the corpus of the Zentralblatt MATH has already been conducted [2]. The authors examined the text contained in the title and the abstract of the mathematical articles and also performed advanced methods to prepare the formulae contained in the abstracts for automated classification. Figure 1 shows the confusion matrix of the resulting classifiers for the 63 top level classes of the MSC. The experiments show that only a few top level MSC categories can be annotated with acceptable quality while most
categories have a high confusion among each other. Consequently, the experiments trying to specify the definite MSC class on the lowest level of the MSC performed even worse.

Of course the problem that manually created taxonomies that have been manually maintained over centuries cannot be annotated automatically out-of-the-box with high precision is not a problem that is restricted to mathematics. This indicates that digital libraries will very soon not be able to retain the high quality semantic annotations as manual indexing becomes more and more infeasible and automatic indexing is too inaccurate. It therefore seems that digital libraries will in future be forced to either rely on low quality automatic annotations or to restrict the semantic annotations to the most important journals.

In [3] the authors present an approach that enabled context searches for chemical entities, despite the documents’ lack of suitable context annotations for the domain. The authors presented a similarity measure using cross-domain knowledge gathered from Wikipedia. Even though the quality of their annotations was not overwhelming, the annotations were still useful to specify and personalize a user’s search context. For digital libraries this means that low quality annotations are still better than no annotations at all to provide personalized context sensitive information retrieval. However, it is questionable if this matches with the high demands of quality that are associated with digital libraries.

3. **Focused Digital Libraries**

It stands to reason that sooner or later digital libraries need the assistance of machine learning approaches to cope with the growing amount of digitally available information. Unfortunately, all evidences show that automating the annotation work of the editors in digital libraries will not work out-of-the-box. The main reason for this problem is that machine learning algorithms were first regarded at the very end of a long evolution of semantic metadata annotation and were not regarded from the start.

To cope with the growing flood of newly published documents and still be able to scale, the focus of a subject-oriented digital library must be to phase the structured domain knowledge in form of ontologies, taxonomies or controlled vocabularies with machine learning technologies. This means that structured domain knowledge must be designed in a way that it can be annotated using machine leaning technologies but still domain experts must be able to understand the semantics of the automatically generated annotations. Finding and maintaining such structured domain knowledge is certainly not an easy task but it is a task that does not become harder with increasing amount of documents.

Below, we will give a short introduction in related research on topical information extraction as well as an overview of current trends in that topic.

3.1. **Subject-centered Information Extraction**

Every scientific domain has developed a certain dialect where natural language is mixed with domain specific terminology. The ability to understand and extract this terminology reveals great opportunities for the creation of value-added services for subject-centered retrieval systems.

In the area of mathematic, for instance it is natural to use formulae in running text. The formulae in mathematical texts are used as arbitrary parts of speech like adjectives, objects, subjects or whole sub sentences. The authors of [4] e.g. describe the contribution of their work as follows:

"[...] We prove in particular that if \( f(x) = ax^n + bx^m \) permutes \( \mathbb{F}_q \), then \( p - 1 \leq (d - 1)^d \), where \( d = \gcd(n - m, p - 1) \), and that this bound of \( p \), in terms of \( d \) only, is sharp. [...]"

We see here that the formula \( f(x) \) is used as a subject and \( \mathbb{F}_q \) as an object in the first sentence. The authors further state that "if \( f(x) \) permutes \( \mathbb{F}_q \), then \( p - 1 \leq (d - 1)d \) using the inequality \( p - 1 \leq (d - 1)d \) as a conditional sub sentence. The same happens with the term \( n > m > 0 \) to express that \( n \) is greater than \( m \) and both \( n \) and \( m \) are greater than 0. It is also striking that strong domain specific notations were used when constraining the range of certain variables with a sub sentence starting with a leading "where". A parser that understands formulae together with the context around the formulae would also be able to distinguish between an article that uses a certain formula or proves a certain formula or under what condition an article proves a certain formula.

In [5] the authors propose a graph based approach for mathematical knowledge management. The proposed theory graphs approach as a representation paradigm for mathematical knowledge that allowed making the modular and highly networked structure of mathematics explicit and therefore machine-actionable. Using this approach for example in digital mathematic libraries, it reveals the potential for computer-supported or even
automatic representation, cataloging, retrieval, refactoring, plausibilization, and in some cases even application of mathematical knowledge.

To give another example let’s look into the field of chemistry. If an organic chemist describes a synthesis procedure it may read as follows [6]:

“[…] 5-Cyclobutyl-2,3-dihydro-[1H]-2-benzazepine 82: Potassium carbonate (0.63g, 4.56mmol) and thiophenol (0.19g, 1.69mmol) were added to the 2-nitrobenzene sulfonamide 50 (0.50g, 1.302mmol) in N, N-dimethylformamide (33cm³) at room temperature and the mixture was stirred for 16h. Deionised water (50cm³) was added and the aqueous phase was extracted with ethyl acetate (5×50cm³). The organic extracts were dried (MgSO₄) and concentrated under reduced pressure to give the title compound 82 (0.259g, 1.302mmol, ca. 100%) as an oil used without further purification.

Again, very strong domain specific terminology and conventions can be observed. Procedure descriptions in the field of organic chemistry always show a high amount of domain specific terminology (like X was added to Y at room temperature, X was extracted with Y, etc.). A system that is able to transform above example into a structured form would also allow building a system to compare and search for synthesis procedures. Of course, this is only a very specific example - an overview summarizing state of the art methods for information extraction used in chemistry is described in [7].

These two examples illustrate what benefits domain specific natural language processing can provide compared to general purpose approaches. Both examples give much opportunity for the application and development of automated information extraction methods like e.g. for the extraction of named entities, domain specific synonym detection or disambiguation. Scientific papers are also a promising target for the extraction or the computer-supported generation of structured knowledge, like ontologies, taxonomies or thesauri.

Currently, structured domain knowledge is almost exclusively created completely manually. Having in mind that this structured knowledge must cover all aspects in the current research in the domain, it becomes clear that the creation and maintenance of such knowledge is related to a huge amount of effort and high personnel costs. Additionally, to ensure completeness, soundness and actuality is certainly not an easy task. Therefore, research trying to generate domain knowledge automatically or at least semi-automatically, became a popular area of research in computer science. The general idea to generate domain knowledge is to extract Salient Terms and relationships out of the document corpus of interest and represent them as ontological knowledge. Basic technologies for this task range from Natural Language Processing [8], [9] over probabilistic language models [10], [11] to statistical approaches adapted from the area of Information Retrieval [12]. In recent years, research has also applied on so called Folksonomies [13] where Data Mining Techniques are applied on metadata provided by users (like e.g. user tags). These so called “Light-Weight-Ontologies” can, for example, be used as a compromise between carefully created ontological knowledge and a loose collection of metadata.

Another point to consider is what sources should be regarded for the annotation of metadata. Of course, the obvious source is the text belonging to the document itself but the whole picture of a document’s context is naturally not restricted to the document’s text. Popular sources to enrich the amount of information for a document are citation or author networks. The properties of citation networks are known very well and have been studied over many years [14], [15]. Citation/author networks can for instance be used to boost the quality of annotations [16] or to rank publications or authors based on the structure of the citation graph [17], [18].

Another popular, newly occurring source of metadata is the Social web. Recently, analyzing this data to estimate the impact and quality of scholarly publications gets more and more popular under the term of altmetrics [19], [20]. These metrics try to reflect activity in social media services with the purpose of gathering scholarly impact besides the traditional citation based metrics. Tracking these activities, it is possible to monitor the manner in which scholarly documents are disseminated and discussed in a narrow time frame [21]. The role of social media in scholarly communication has been investigated in several studies including their use in dissemination [22], conference chatter [19], science popularization [23], and promotion of scholarly products [24]. In addition, several tools have been introduced to facilitate the use of Altmetrics, e.g. PlumX, ImpactStory, Altmetrics and Scholarometer [25].

As the target of this article is only to give an intuition of the capability of topical digital libraries, we will not go into details of information extraction here. For more information on this topic we therefore refer to [26].

3.2. Creating Useful Rankings

The ultimate goal of a digital library (as for every information retrieval system) is to answer the users’ information needs with plausible rankings. As already stated, to succeed in this
task domain specific information extraction is required to represent the knowledge and semantics of a domain and to annotate documents and entities in the domain, respectively. The basis of building such a ranking is a similarity measure.

However, computing a similarity of two entities based on given semantic annotation is also not an easy task. In chemistry there exist a whole stack of similarity measures between substances in use, like the Tanimoto measure, the Russel-Rao dissimilarity, the Yule distance and much more. The question what similarity measure is best suited for chemists is not easy to answer. In [27] the authors compared and evaluated different similarity measures and tried to find out if the metrics are redundant or if they represent different concepts of similarity. An important observation was that that the similarity measures show almost no correlation, meaning that they certainly represent different aspects or definitions similarity. It was also shown that the assessment of the quality of a similarity measure is highly dependent on the individual chemist, suggesting that different similarity measures correspond to different ways of chemists’ perception of similarity. The authors also showed that by using feedback provided by the users the quality of rankings could be improved significantly after only very few feedback cycles. This observation shows that subject-oriented libraries can improve the search quality greatly by using domain specific structured knowledge. For text search this improvement might still be a nice-to-have feature, but for domain specific entities like e.g. molecules the annotation of domain specific structural knowledge is mandatory to provide a plausible ranking.

4. Conclusion

This article aims to raise the awareness of the necessity of subject-oriented digital libraries as opposed to “one size fits all”-style knowledge repositories. As textual structures, conventions, structured knowledge as well as requirements to a retrieval system differ tremendously among different domains, subject-oriented digital libraries can produce great benefit for experts in the respective domains.

An essential task to provide domain specific value-added services is the annotation of structured domain specific knowledge like e.g. keywords from ontologies, taxonomies or controlled vocabularies. However, the annotation of these keywords as is it performed today will not work out in the future. Also, automated indexing approaches where state-of-the-art machine learning algorithms are “plugged” on top of a long evolution of manual indexing work are proven to not work properly. We therefore claim that in order to retain high quality semantic annotations and, on the other hand, still are able to scale, it is necessary to develop structured domain knowledge that sufficiently reflects the semantic of the domain but is able to be annotated automatically by using state-of-the-art machine learning technology.

We also gave an insight what topical digital libraries can target. We showed several examples of domain specific texts and discussed what kind of information extraction methods can be applied to the texts that are restricted to that particular domain. We also looked into different sources to obtain metadata; especially we introduced the newly occurring idea of altmetrics that utilizes data from the social web to assess the impact or quality of a publication.

References


Learning Management Systems
- A Need for Specialized Systems

Sharif, Nauman; Afzal, Tanvir; Helic, Denis

Abstract - eLearning is considered a multibillion industry today having a rapid growth. There are many general eLearning systems available which offer the same architecture and services for all domains of sciences. These general Learning Management Systems are reviewed in this paper. Furthermore, some of the important research efforts are highlighted which are useful in a limited learning domain. As a consequence, we show that, due to massive expansion in eLearning industry, there is a need to provide subject specially designed eLearning systems to meet the individual needs of the learner rather than generic eLearning systems supposedly applicable to all disciplines. Such specialized eLearning systems should particularly emphasize active and authentic learning.

Index Terms: Specialized eLearning, General eLearning systems, LMS, E-learning recommendation

1. INTRODUCTION

Recent years have seen a huge proliferation in the eLearning industry. “In 2011, it was estimated that about $35.6 billion was spent on self-paced eLearning across the globe. Today, eLearning is a $56.2 billion industry, and it’s going to double by 2015.” [3]

As a result, traditional learning management systems (LMS) are faced with the challenge to provide not only a true alternative to brick and mortar class room environment but to provide a more enriched learning experience.

Scientists have started to question the effectiveness of traditional eLearning systems. Although eLearning systems have been a great success in some cases, investigations into whether eLearning has been successful in bringing about the long anticipated paradigm shift have shed some doubt at general approaches. [4]

In this paper, the authors have investigated available classical eLearning systems as well as research proposals and innovations suggested by the scientific community to support active and authentic learning.

In section 2 general eLearning systems are critically examined and how they fall short in fulfilling the need of a diversity of learners. In the following section scientific community contributions are investigated and possible ways are identified which can enhance the learning experience during the learning process. Section 4 highlights the need to utilize scientific community contributions to create a specialized eLearning system.

2. GENERAL LEARNING MANAGEMENT SYSTEM (LMS)

In a traditional classroom environment an instructor is able to obtain feedback on student learning experiences in one-to-one interactions with students. Instructors can assess the learner needs by various means. For example, a learner’s previous learning experience can provide useful information about his/her learning style. This feedback mechanism allows the instructor to recommend appropriate learning resources and tasks to support the individual’s learning experience. [1]

With the evolution of the Web, scientists focused on computer based eLearning systems i.e. learning electronically without physically appearing in the classroom [28]

There are varieties of LMS available in the market which support online learning by creating course material, designing student assessments such as quizzes, assignments etc. and provide online forums and blogs for peer learning. Examples of the most popular commercial LMS are blackboard/ WebCT, JoomlaLMS [17]. Similar free LMSs are Moodle, Sakai, and Docebo [18].

These LMSs are largely seen as a mean to teach many students in an adequate manner. In traditional LMSs educational material is planned and designed by an educational institute and the instructors, whereas the learner is expected to interact with a predefined pedagogical process. A generic solution is used across different domains.
of education such as computers science, math, biological sciences etc.

Educators select the content for learning. In other words, knowledge is presented to the learner without considering the individual need of the learner. [2] A generic solution is applied across culturally and linguistically distinct learners. Hence we can call them ‘general LMS’.

Let us look at an example of Moodle, an open source Course Management System. Moodle is offering over 1.8 million courses, used by 1.7 million teachers in 270 different countries [30]. It is widely used by different universities, colleges and training institutes to provide online course contents through web. Moodle offers instructors, students and educational institution the following features: Files download, assignment submission, online quiz, Moodle instant messages, online calendar, online news and announcement (College and course level), discussion forum, Wiki and grading. [7]. Moodle can basically be seen as a course management tool facilitating the instructor to create and manage online courses.

Although Moodle has some great success stories to share, whether it provides an effective personalized learning experience across different disciplines is questionable.

In an eLearning environment learners can discover the relevant knowledge by queries or by following navigational pattern set by the education provider. Learners are often confused to the amount of information that is presented to them.

Contents need to be presented in a supervised manner at the right time of learning i.e. when contents are most appropriate for the learners rather than learners selecting irrelevant content for the current learning context [31].

Learning management system should be more centered towards individual learner’s need rather than mainly supporting the teacher as a course management tool. Technology should be seen as assisting individual learning experience rather than manipulating the static learning contents. Many of these generic LMSs provide a static approach across the different levels of learners assuming all learners are equally skilled to learn a particular subject matter. Hence a generic pedagogy is adopted to meet the needs of many learners without considering their learning diversity.

Let us take an example of learning a new topic in a general LMS environment. There may be some learners who require further elaboration only concerning this new concept. These learners simply want a few recommendations on the most relevant material for the current learning context. However, general LMS do not provide recommendation from the web to support the current learning context of the learner [31].

In general, LMSs environment learners are left to seek the help from the web. When the learner searches the information in the web a huge amount of information is presented to the learner. The learner is faced with the challenge to find the most relevant information according to his/her learning context. This can be a very tedious task! The learner finds real difficulty in creating a knowledge map tying together different concepts.

As a result, the learner is left in disarray and disengages from the learning process. However, in physical class room environment, the instructor can easily recommend few of the most relevant resources from diversified references.

Some research projects have attempted to propose innovative techniques for providing recommendations from external sources, i.e. from outside of eLearning environment. However, these systems have been limited to particular subject areas. In the next section, a comprehensive review of some of these intelligent eLearning systems is presented.

3. ADAPTIVE AND INTELLIGENT ELEARNING

The emergence of web2.0 has brought about new innovations in the field of eLearning environments. [33]

“The Web 2.0 revolution has peddled the promise of bringing more truth to more people, more depth of information, more global perspective and more unbiased opinion from dispassionate observers”[19]

Today’s ‘digital native’ learner has the powerful tool of social networks where the learner can independently create, publish and redistribute contents. This learner finds general LMS structure inflexible as compared to learner-centered approach of Web 2.0 services. [20]

The scientific community is actively engaged to make the learning experience increasingly productive and fruitful across distinct needs of the learner. This has led to a paradigm shift from teacher centered to learner centered education [21].

Locating recommendations in external sources for learners has remained an active area of research [22]. Recommendations in eLearning are somehow different than in other domains. For example, one has to keep in mind the learning context of learners, their knowledge history, and their profiles. Furthermore, recommended resources need to be judged whether they are a new concept or continuation of the existing knowledge [12].
Recommender systems for eLearning domain have been proposed and evaluated by scientific community in small settings. For example, Kumaran & Sankar proposed such a recommender system which uses learner information and domain knowledge (computer science). They employed semantic nets for modeling the learner profiles and the learning contents to provide personalized recommendations [23]. However, this system has been evaluated for only one particular course.

Similarly, there is the emergence of adaptive and intelligent systems where learners can create personalized learning environment rather than technology providing static learning contents. Learners should be able to create, manage and organize the knowledge according to their personal knowledge management capabilities. Learners’ previous learning experience and current context can be used to provide personalized and adaptive learning experience.

These adaptive and intelligent systems are the joint venture of intelligent tutoring and hypermedia systems (AHS). Some examples of domain specific ITS are SQL-Tutor, German Tutor, ActiveMath, VC-Prolog-Tutor, similarly examples of AHS are AHA, InterBook, KBS-Hyperbook WebCOBALT. [8]

Adaptive and intelligent eLearning system can be obtained by modeling a domain (using e.g. some ontology structure) or using pedagogical datasets (sets of designed problems and their solutions), or data about user interaction and learner’s model [9].

In order to provide personalized learning experience, various data mining techniques have been used. For example Koutheair et al recommends a system which mines users’ web usage and learning materials in order to predict personalized contents to an active learner [11].

A personalized content recommender system proposed by Lu et al applied fuzzy matching to determine relationships between learning needs and list of learning contents [24]. They tested the system on specific courses. However, the scalability of such a system needs to be verified on diversified courses and domains.

Tane et al. applied text clustering and mining rules to arrange documents as per their topics and likeness [25]. Dwivedi et al. used weighted hybrid scheme to recommend relevant learning contents to the learner by modeling learning style and the knowledge using collaborating filtering technique [26]. Such systems, when integrated in general eLearning system may not produce accurate results, since learners from different educational disciplines will have different learning styles. Learners’ collaborative data such as: co-downloads, co-views, click streams may not be relevant to all.

In addition to the mentioned recommender systems, social networks have attracted learners to build knowledge communities and recommend knowledge resources to each other [16]. Wiki-Learnia is an example of such system where knowledge is extracted and distributed from social networks such as Facebook, Twitter and YouTube. Authors have populated subject ontologies. However, providing such features in a general LMS is challenging task. Ontologies from all educational domains need to be built and updated accordingly. [13].

4. A NEED FOR SPECIALIZED E-LEARNING SYSTEM

The scientific community indicates that there is a huge gap between the general LMS and needs of ‘Digital Native’ learners. General LMS were designed by ‘Digital immigrants’ who have failed to cater to the ever growing demand of ‘Digital Native’ learners [5].

Instead, specialized eLearning systems should be provided to learners which could recommend relevant contents in the current learning context. It should provide flexibility to learners. A Learner should be able to discover, organize and share information in a locally meaningful fashion which is globally accessible. [15]

A general eLearning system provides a generic profile for every learner. Static contents are presented to all learners who may have different skill levels. However, a specialized system can dynamically build learners’ profile and delivers appropriate contents at right time to support individual learning. Dynamic profiles can be built by considering learners’ recent navigational history. Subsequently, content similarity and dissimilarity can be mapped to learners’ dynamic profile [6].

As it was investigated in section 3, the research community has provided a significant contribution towards active and authentic learning. A variety of recommendation tools are successfully recommending appropriate contents in a limited educational discipline. There is a dire need of specialized eLearning system which can model the entire educational domain. In this manner specialized eLearning systems can provide better recommendations according to personalized needs of learners. Specialized eLearning systems can be semantically more affluent and provide better diagnostic analysis than general LMS [10].

5. CONCLUSION

This paper investigates the limitation of general Learning Management Systems (LMS) and concludes that even though general LMS
provides good course management features, such general LMS do not truly meet the challenging needs of ‘Digital Native’ learners.

The scientific community has provided significant contributions towards true learning experience. These innovative efforts are comprehensively examined and it was highlighted that general LMS have failed to incorporate such important eLearning features in their systems. Finally, the paper highlights that there is a need to build specialized eLearning system for each educational domain. Such specialized eLearning systems can provide true eLearning experience to learners according to their particular needs, context, profiles, histories, collaborations etc.

References


[29] https://moodle.org/stats/ (accessed on 10th April 2014)


An Educational Framework for Content Sharing

Baumgartner, Peter; Zobel, Annett

Abstract - Open Educational Resources (OER) did not show the positive impact many educators and politicians had expected. Instead of building up repositories organized just by content criteria the paper proposes a community platform (edu-sharing.net) supported by a search engine triggered by criteria taken from an educational taxonomy of teaching methods.

Index Terms: Open educational resources, Educational taxonomy, Educational pattern, Content sharing, Reusability.

1. INTRODUCTION

The use of Open Educational Resources is still very disappointing. Together with Gert Kortemeyer we are posing the question: “Why do the vast majority of higher education venues still depend on expensive paper texts, while most of the world’s knowledge is available for free online? Why do educators not embrace the plethora of open digital educational libraries and repositories?” [1]

We have analyzed different barriers to overcome for using OER. All of the problems are related to educational issues not to technical limitations [2]:

(i) Educational requirements: Finding the right resource is a question of economy of scale. Teachers are not looking for educational material as such but for an object with many detailed educational characteristics. This desired list of qualities is linked with the “and” operator and is therefore limiting the search result with every additional property.

(ii) Educational metadata: In spite of sophisticated federated search engines and well known huge content portals (e.g. [3]–[6]) we are still lacking a sound educational taxonomy which teachers use and understand. Learning Object Metadata (LOM) is not sophisticated enough to fulfil practical educational needs of teachers.

(iii) Educational culture: The “not invented here” syndrome and the well-known problem that learning objects created for a limited personal usage have to undergo still a long and cumbersome process to make them fool proof for every possible standard situations are two sides of the same coin and limiting the use of OER.

(iv) Educational quality assurance: Who has the necessary qualification and authority? This is not only a question of competence but in a participatory community model also a question of regulatory procedure and power relations. What kind of agreed and fast procedure is to follow? The blind peer review as the traditional model of quality assurance in science is not only far too slow but also seems inadequate in an open community model of fine grained different needs and diverse interest/target groups committed to a variety of educational models and approaches.

With the edu-sharing network [7] we try to overcome these problems and limitations. This paper explains the rationale for our approach, the implementation so far and future plans.

2. THE IMPORTANCE OF A COMMUNITY PORTAL FOR TRUST BUILDING

We know that material that is offered via Internet only by real names – or worse – by nicknames is not sufficient for trust. Confidence building is a cumbersome process, which has to be regarded from two sides: from the motivation of the supplier of the resource and from the interests of the user of the resource.

Let us start with the motivation of the supplier of the resource: Lacking direct financial compensation in open content portals we have to look for different motivational reasons for passing on material that one has created or adapted.

The hope to get other material in exchange is generally soon disappointed. We know that in the Internet culture there is no symmetry between supply and needs. Anderson has convincingly shown that there is a “long tail” of supply [8], meaning that very few products generate the main income – or in our context – satisfy the majority of needs. From a financial viewpoint this speciality of the Internet economy makes perfectly sense. As for digital material there is almost no storage cost [9] so that every single buy of a product – even if it happens not

Manuscript received April 15, 2014.
Peter Baumgartner http: is with The Danube University Krems Austria. E-mail: peter.baumgartner@donau-uni.ac.at
Annett Zobel is with. Edu-sharing.net association, Germany. E-mail: annett.zobel@Edu-sharing.net

48
very often – generates real revenue. It is therefore beneficial to provide all kind of dead articles or slow sellers. But this strategy does not work in a pure exchange economy where money as general change agent is missing: In that case we need either a reciprocal match between needs and quality of the exchange object between supplier and client (which is very seldom the case) or we would need a non-financial neutral exchange mechanism. In community networks normally this non-financial exchange mechanism is reputation. Where the objects are already available e.g. ready at hand and finished (like music files) there are no excessive additional costs except for uploading time and – if there is no flat rate – communication or connections costs. But the first option (finished for sharing) is generally for educational material not the case. Learning objects have not only to be described with metadata, but also to be supervised and explained in contextual details in order to get helpful for a broad public. This “last mileage” of this preparation work for dissemination produces high additional costs. Especially if one of the non-financial exchange values is reputation, there is a big barrier to pass on material that has not incorporated this additional work.

Different approaches have already been developed to overcome the mentioned problem:

(i) A quality assurance procedure by the portal externalizes these additional costs and is only valuable if there are many committed participants working without money or if there are financial funds available. Another disadvantage of this approach is the vaporization of the individual reputational value as many different persons are working on the same objects.

(ii) To prevent jumping on the bandwagon and just use material of other people without giving anything back the portal may set up rules to follow. For instance one has to deliver material or other services (giving feedback, evaluating material, writing reviews etc.) in order to get the right to download material. Mostly these kinds of regulations are restrictions and barriers to build up a portal community very fast. This is especially a draw-back for new initiatives as a critical minimum on member participation is necessary for a useful exchange and therefore for a functional content sharing network.

Another way to overcome this obstacle of additional disseminations costs and of the critical mass of participation is to build up trust by getting to know someone personally. If I know someone personally – let us say a friend of mine or a colleague of my department – then there is a common understanding and mutual reliance built up over time. It is helpful but not necessary that we meet in person. We could provide confidence through different interactions like exchanging our views on a certain subject via emails, posting photographs or other personal material, visiting and commenting articles of our weblogs or using one of the different functions of social software incorporated into the portal, etc.

It is one very useful side effect of this approach that (still) small member participation is not an obstacle and has to overcome but is a helpful condition of its own for the trust building processes. Small is beautiful in this case: One personal acquaintance with whom I will share content regularly is more effective than a huge portal with millions of possible assets which I have to search, filter, sort out and try to use on my own without contextual experiences of the producer or developer.

3. **SHARING NOT ONLY CONTENT BUT EDUCATIONAL CONCEPTS TOO**

In principle we can distinguish two different approaches of providing and using OER: Repositories with educational content also known as Reusable Learning Objects (RLOs) [10]–[14] and open access to complete courses like the MIT initiative OpenCourseWare (OCW) or courses as (by-) product from some Massive Open Online Courses (MOOC) providers [15]–[19].

Both sharing strategies do have disadvantages: In order to maximize reusability of learning objects the material has to be free of context. This devalues the content from the educational perspective as material for high quality teaching has to be adapted for specific circumstances like learning goals, previous knowledge of the learner, available other resources, time frame, personal learning style, teaching method etc. This means that learning objects have to be adapted in order to get integrated well into the planned course material, a problem we have reported and analyzed several times [2], [20]–[24].

The other way is to use complete courses that have already applied and integrated all the above-mentioned educational assumptions. But here we also have to face different challenges: Besides the problem that one has to give up to a certain degree his/her individual teaching style we also have to link and cross reference the content of one course to other courses or
modules, meaning that the problems of RLOs reappears at a higher curriculum level again.

The MIT with its long dated experience from their OpenCourseWare initiative which started already October 2002 has launched the MIT Core Concept Catalogue (MC3). MC3 is an academic data service to “manage and share information about the curricular topics, learning goals, and related content within and across disciplines and subjects” [25].

The aim of this relatively new service is to integrate different parts of content (courses) under educational premises: “Just publishing content is not enough. We must find simple and scalable ways to expose the underlying concepts, learning goals and their relationships so that educational content can be more easily aligned, aggregated and re-used across departmental and curricular boundaries.” [26]

4. THE EDU-SHARING INITIATIVE AS A NEW AND ALTERNATIVE APPROACH

From our point of view even the very advanced MIT academic data service is limited in two ways:

(i) MC3 is constrained to the MIT community and their published course content.
(ii) MC3 is a service that works as an additional data layer that is not integrated into the educational content production.

What is necessary is a development tool, which combines the planning of the content sequences with the planning of the learning activities (= educational methods). Only the visible and therefore modifiable integration of content objects with educational scenarios objects will provide teachers with lesson-plans they can adapt and elaborate to meet their own specific needs.

In the edu-sharing.net initiative educational organizations and users have full control about their contents because the edu-sharing repositories are installed within the educational organization. The repositories can be connected to the edu-sharing network if the organization wants to allow content sharing with specific external persons or organizations.

Edu-sharing users perceive the repository as a network drive, which can be connected to their Windows Explorer or Mac Finder. This can be thought of as a kind of Dropbox [27] for education. User can drag and drop contents between local file folders and edu-sharing network folders or they can save content out of their authoring application to the network drive or they can insert edu-sharing content from the network drive within the current document (e.g. a Word file). Furthermore edu-sharing provides a metadata editor and a powerful content search.

All content can be easily used within applications for learning such as Learning Management Systems (LMS) like Moodle [28] or other collaborative tools that can be used for learning (e.g. wiki) because the repositories are usually deeply integrated within the IT infrastructures of the organization. The main target is to reduce media breaks through cloud services even though people work in different organizations with different LMSes.

5. FUTURE PLANS WITH EDU-SHARING

The next step is to build up a community portal between the distributed edu-sharing repositories and to add the possibility to share not only content but also tools and edu-patterns.

Our main thesis is: To improve the reusability of learning material we would need an editor for lesson plans where we can search and integrate content, tool, and educational patterns.

Quite a bit of work has already been invested to facilitate the design of digital lesson-plans. To support a wide range of different educational strategies the Open University of the Netherlands (OUNL) developed IMS Learning Design (LD) 2003 as a standard to describe the specifics of different pedagogical approaches for online learning [29], [30]. But this standard is very complex so that we are still lacking implementation on its most advanced level. The Learning Activity Management System (LAMS) – inspired by the ideas of LD – is a more practical approach. It creates digital lesson plans – collaborative learning activities – that can be run online with students and shared among teachers and between different learning management systems (LMS) as well. LAMS uses a visual authoring environment for creating learning paths as a sequence of different learning activities.

These are already some tools available for digital lesson planning [31]–[34]. But what is still lacking are editors which search for content, tools and educational patterns and are able to integrate these different ingredients to one lesson plan which is educational sound.

The following example with the “gallery method” [35] – a complex educational pattern and creativity technique – will demonstrate our plans for the first integrative steps: linking content with educational patterns:

In the traditional face-to-face situation the gallery method consists of five consecutive steps but these phases have to be conceptualized in eLearning in a complete different way. In real classroom settings group
building is a precondition for this method, which normally does not need an explanation, as anybody knows how to do it. As a consequence the concrete procedure in teaching guides is not explained even not mentioned as a separate step. In virtual scenarios however one has to put attention on group building processes because it is not easy to coordinate the necessary activities online. Special tools have to be provided to support this activity in virtual settings.

In the next implementation of edu-sharing developers of lesson plans (teachers) will have the possibility not only to search for content and tools but also for educational patterns. If they decide to use a special edu-pattern then they will be guided to the adequate tools for these special educational scenarios to facilitate the design of the whole learning situation. Edu-sharing will therefore bridge the gap between tools (= lower level activities) and educational pattern (= higher level activities). This is important as teachers are used from the face-to-face scenarios to think in higher-level activities.

Instead of just offering some educational pattern edu-sharing.net will implement a comprehensive educational taxonomy of teaching methods drawn from a published book where they are discussed systematically [36]. They will be collected and presented as a pattern language [37], [38]. The idea of a special pattern description format has spread from architectural design to the design of object-oriented programs [39], [40], web design and human-computer interface design [41], [42] to other non-technical areas like educational design of teaching/learning scenarios [43]–[45]. Until now there is no systematic implementation of an educational pattern strategy in eLearning. Edu-sharing.net will be the first experiment in this direction. The future and a thorough evaluation of use cases will show if this approach is able to overcome some of the problems we reported in this paper.

6. SUMMARY AND CONCLUSION

We believe that the not very enthusiastic use of open content by teachers is partly caused by missing educational scenarios suitable for online lesson planning. It is not the content alone that generates a learning opportunity but a specified learning activity (e.g. teaching method) as well.

We therefore advocate special repositories where teacher can search not only for content but also for educational patterns to implement learning activities appropriately in their digital lesson plans. The presented network edu-sharing.net is a first step to integrate educational content with educational patterns to facilitate online learning.

REFERENCES


AUTHORS

Peter Baumgartner http://peter.baumgartner.name is Professor of Technology Enhanced Learning at the Danube University Krems (http://www.donau-uni.ac.at), Austria.

Annett Zobel is computer scientist and firm owner of metaVentis (http://www.metaventis.com) in Jena/Germany. Currently she is the managing director of the board of the edu-sharing.net association.
Heritage Portals and Heritage Mining: Synergizing Data and Image Mining under Uncertainty Constraints

Milutinovic, V.; Salom, J.; Jelisavcic, V.; Filipovic, V.; Mihajlovic, A.; Ognjanovic, Z.; Markovic, Z.; Kos, A.; Tomazic, S.; Maurer, H.

Abstract - This paper first discusses the major requirements that a heritage portal has to satisfy. In this context, the term heritage portal refers to portals for digital preservation of national heritage. The listed requirements are divided into two categories, primary and secondary. Second, this paper portrays the implementations of the above requirements in the context of Serbian Forum. Third, the paper introduces a novel approach to data mining of historical data, the heritage mining approach, for which heritage portals represent the enabler technology. A generalized algorithm for heritage mining is given with relevant details and with one relevant example.

Index Terms: Data mining, Digital preservation of national heritage, Heritage.

1. INTRODUCTION

Heritage portals [Korica2012] are specific in many aspects. They include books, documents, historical, and geographical information of encyclopedic character, but also music, still images, and moving images. All this represents a BigData collection with lots of hidden knowledge. Mining the history helps predict the future. However, mining from heritage portals is extremely difficult for a number of reasons:

(a) It is multi-dimensional (text, music, images), and therefore complex.
(b) This complexity is made even higher, because of the need for different dimensions to synergize.

(c) In addition to the complexity, one has to cope with uncertainty constraints, arising from the fact that data could have been corrupted, intentionally or unintentionally.

We will review the types of data corruption after we review the basic characteristics of heritage portals; because the heritage portals represent the enabler technology for implementation of heritage mining (heritage mining is best done via cross-correlation of data from heritage portals of different nations). Consequently, characteristics of heritage portals directly influence the types of data corruption.

2. HERITAGE PORTALS

Heritage portals must pose specific characteristics. These can be divided into two major categories:

(a) Primary characteristics,
(b) Secondary characteristics.

Primary characteristics of heritage portals are:

• The content is controlled by national institutions established by local governments, with an obligation to worry about digital preservation of national heritage. This is different from the Wikipedia case, where the content is provided by individuals and is controlled by Wikipedia.

• The content is protected by a plethora of legal regulatory. Not only by Creative Commons (CC). Each particular document can be protected using a different type of legal regulatory. For example, one document can be copyright-free, while another one can be copyright-protected.

• The content must include titles with short commentaries translated into foreign languages using the culture-oriented translation approach. This means that translations into foreign languages must include hypertext, which makes it semantically understandable to an average reader of the targeted language who grew up in his/her
native culture. For example, if a sentence is translated from Japanese to English, and the term “shogun” is used, a hypertext link must be included. Something like that does not exist in Wikipedia, and Google Translate will not produce anything similar.

- The content must be presented in a way that favors quality (and ranking), not quantity (and chaos). For example, the top N artifacts from each contributing institution must be ranked, which enables a viewer in a hurry to see quickly the top N contribution of each contributing institution. Something like this does not exist in Wikipedia.

Primary characteristics will always represent differences between Wikipedia on one side and heritage portals on the other side.

One heritage portal that possesses all the above characteristics is Serbia Forum (www.serbia-forum.org). It originated from Austria Forum (www.austria-forum.org). Another member of the family is the heritage portal of Academia Europaea (www.ae-info.org).

Secondary characteristics are also important. They refer to functionalities that enable more comfortable viewing of heritage portal contents. Secondary characteristics are not unique to heritage portals. If they exist in heritage portals and not in Wikipedia, two future scenarios are possible:

(a) If they are innovative and good, Wikipedia will soon acquire them, and differences will disappear;
(b) If they are innovative but bad, Wikipedia will not acquire them, which means that the differences will remain to exist, but will be irrelevant.

Lacking some of the above characteristics is bringing the decline to Wikipedia [Simonite2013]. Talking about Austria Forum and Serbia Forum, important secondary characteristics include:

- Information must be semantically searchable, by metadata:

  We should be able to search for someone whose name we do not know, but we know he was born in Vienna (Wien), was working in physics (Physik), and died in Italy (Italien). The search should find Ludwig Boltzmann.

- The author must be visible: The source of the text must always be known.

Either the name of the author including CV, or the book/archive it comes from, should be visible. The same visibility-related data must be kept on all updates.

- Evolution must be tracked: All versions of the archived item must be reachable,

  All versions - from the first update to the last (the present state) should be provided to an inquiry. Note that updating could be dangerous if it destroys previous information and, hence, history.

- Existence of original commodities: Books are not just a source of information.

  Web books behave like printed books, but also offer some new functions, links, bookmarks, etc.
  Books should be able to turn into social networks, by enabling discussions, comments, etc.

  These characteristics will exist, also, in a number of other national heritage portals currently under construction. There is an ongoing Europe-wide development effort to be reported in a follow up paper.

3. COMPONENTS OF THE SYNERGY

Types of mining of interest for heritage portals are listed below. Each type has to be applied independently of others and after that, results have to be compared for consistency. The methods include:

a) DataMining (DM) from general databases (DB),
b) TopicMining (TM) from specific databases,
c) ExpertMining (EM) from the Internet-based Social Networks (SN),
d) PsychologyMining (PM) from the Internet-based SNs (psychological profiles, mindsets, users’ sensibilities etc.),
e) DM from still images (both indexed and non-indexed, which implies utilization of image understanding tools),
f) DM from moving images (both indexed and non-indexed),
g) DM from Wireless Sensor Networks (WSNs), which is of interest for verification of geo-historical data, and
h) DM from the Internet of Things (IoT), which is important for observing the present environment.

All the above methods are elaborated in [Milutinovic2013]. The uncertainties that the
methods may encounter are:

(a) Information is missing (for example an event happened and was not recoded, or was recorded and, later, removed). This type of problem is healed using an imputation method [Mihajlovic2013].
(b) Information exists and should not be there (for example, an event had never happened but was re-coded). This type of problem is healed using an amputation method.
(c) Information was forged (for example an event has happened but the recoding is false). This type of problem is healed using a mutation method.

A pseudo code of the generalized heritage mining is presented next:

1. Collect knowledge in raw form (fill the heritage portal with text, video, audio data...),
   determine the data clusters, and form the average distances between different clusters
   (the stride of the algorithm).
2. Select a point of interest (time period, geo location, person/object of interest...), and declare it the starting point of the extrapolation process;
   check the authenticity of the data involved, using heritage portals of other nations
   (enabler technology).
3. Initiate the parameters:
   maximum search depth (max_depth),
   current search depth (curr_depth), etc...
   This sets the foundation for comparison of data in different portals.
4. Find all related data-points in the current heritage portal (using standard data mining methods);
   a. If no data was found, or too few data points were found, then:
   b. If curr_depth < max_depth: Find all related heritage portals
   c. Repeat the step 4 for all related heritage portals
   d. Go to 5.
   e. For each data-point found, mine related data-points from related heritage portals.
5. Collect all found data-points in step 4 along with their originating heritage portal’s ids.
6. If data is found missing in the starting heritage portal, do imputation using collected data;
7. If a non-related data point is found in the starting heritage portal, do amputation;
8. If inconsistency is found between the collected data-points, do mutation;
   mutation is best done by combining the data from various heritage portals,
   and by finding "truth" somewhere in the middle way between data of various portals.
9. Extrapolate to a future point and derive a prediction.
10. Check the likelihood level of the prediction, be extrapolating into the past;
    if the extrapolation into the past leads to finding an event that supports the prediction,
    the likelihood of the prediction is considered high.

Organized, argumented discussions over the Web using Argument Web platform and tools [Bex2013] or similar, might be of great help in all these three processes (imputation, amputation, and mutation). Information from various heritage portals should be compared for consistency. If an inconsistency is noted, cross-correlation, combined with imputation, amputation, and mutation, may lead to a plausible conclusion [Moskowitz2006].

Generalized algorithms are best understood using an example such as the following one.

4. AN EXAMPLE OF HERITAGE MINING

This example of heritage mining is related to the Serbian history and Serbian forum. In this case, the sequence of presented events is obtained using either imputations or amputations or mutations, or a combination thereof.

In this example, the heritage mining algorithm implies the following steps, as indicated above:

1. To determine the average distance between generations of data sets in a heritage portal, using a specific number system. For example, for the last two centuries the average distance between generations of people in Serbian families was about 27 years.
2. To determine an important event that drastically changes the entropy of the system. For example, one can say that the World War I in Serbia started in 1912 with the first Balkan war. Therefore, for this example, the year 1912 can be determined to be the beginning of the coordinating system.
3. The average distance between two data generations (27 in the above example) should be added, iteratively, from the start of the coordinating system to the present time. At each point of iteration, if an important event is missing, imputation can be done. Where appropriate, amputation is done. If an event exists that is not on the same level of importance as expected, mutation is done. By applying these three rules, the following sequence is obtained for the case of the recent Serbian history:
   - 1912: The first Balkan war started.
   - 1966 = 1939 + 27: The start of the student uprisings, which culminated in 1968.
   - 1993 = 1966 + 27: The start of the culmination of the war in Bosnia.
4. Prediction of the next major political instability in or around Serbia:
- Around 2020 = 1993 + 27: A major political instability predicted.
- Around 2047 = 2020 + 27: Another major political instability predicted.

5. The quality of the above predictions can be determined by going backwards from the beginning of the coordinated system:
- 1885 = 1912 - 27: The war between Serbia and Bulgaria.
- 1858 = 1885 - 27: The dynasty change in Serbia. Aleksandar Karadjordjevic died and Milos Obrenovic was inaugurated.
- 1831 = 1858 - 27: Serbia was given autonomy in religious affairs by the ecumenical patriarchate Concorde. The year 1831 is also the midpoint between Hatt-i-Sharif¹ #1 (1830) and Hatt-i-Sharif #2 (1832); both of them had the major impact on the change of the face of Serbia.
- 1804 = 1831 - 27: The first Serbian uprising against Ottoman Empire - the major event in the national history of the last two centuries.

6. The multiple of the number 27 should also be examined:
- 54 = 2 * 27. Interestingly, the average age of the leaders of each the above mentioned political instability was 54.

One can argue against the above by stating that the mentioned events were taken on purpose to correspond to the pre-chosen years. Arguing that no mayor political instabilities or wars existed in the in-between years may not help. What may help is to take Fourier transform of a vector of data representing the years 1800 to 2014; each data item in this vector has a value in the range of -3 to +3, where the sign tells whether the year related to that data item was positive or negative for the nation, and the magnitude of the data item tells about the intensity of the events representing that year. The expectation is that the peak of the Fourier transform will be around 27.

The explanation of the described phenomena is as follows: In some time period, at some geographical area, some people acquire an energy which results in essential advancements (financial strength, population count, etc.). It is natural that these people would like that the political formalisms (division of revenues, border lines, etc...) follow the essential changes. Of course, those on the essential down-hills do not want the changes in the political formalism. Consequently, political instabilities happen unavoidably. It is logical that the above described political instabilities happen, with smaller or larger amplitudes, once per generation; that is why it is important that the average distance between data generations is established in the first step of the heritage mining algorithm presented above.

5. Conclusion

Following are the main viewpoints of our work presented in this paper.

Firstly, we present heritage portals as main enablers of the heritage mining. By combining data sources from multiple heritage portals, detection, and mining of heritage items of interest could be made feasible.

Secondly, we underline that algorithms for imputation, amputation, and mutation are crucial for mining the heritage portals. Their essence is in cross-correlation of data between various heritage portals, which helps both the starting-point portal and the related portals.

Thirdly, one example is given to shed light on the issues. It is understood, in this example, for each major event, heritage portals of related nations also have to be checked, to determine if the event really happened, and if it happened exactly as described in the starting-point portal.

References


¹ The imperial edict in the 19th century Turkish empire.
Authors of papers are responsible for the contents and layout of their papers.

Reviewers:

**Australia**
- Abramov, Vyacheslav; Monash University
- Beeg, Rezaa; Victoria University
- Bem, Derek; University of Western Sydney
- Betts, Christopher; Pegasus Computing Pty. Ltd.
- Buaya, Rajkumar; The University of Melbourne
- Chapman, Judith; Australian University Limited
- Chen, Yi-Ping; Deakin University
- Hammond, Mark; Flinders University
- Henness, Paul; University of Queensland
- Palmisano, Stephen; University of Wollongong
- Ristic, Branko; Science and Technology Organisation
- Sajianhar, Atul; Deakin University
- Siddhu, Amandeep; University of Technology, Sydney

**Austria**
- Derntl, Michael; University of Vienna
- Hug, Theo; University of Innsbruck
- Lodli, Susanne; Johannes Kepler University Linz
- Stockinger, Heinz; University of Vienna
- Sutter, Matthias; University of Innsbruck

**Brazil**
- Paracho, Aníbal; Universidade Federal Fluminense
- Traina, Agma; University of Sao Paulo
- Traina, Caetano; University of Sao Paulo
- Vicari, Rosa; Federal University of Rio Grande

**Belgium**
- Huang, Ping; European Commission

**Canada**
- Fung, Benjamin; Simon Fraser University
- Grayson, Paul; York University
- Gray, Bette; Alberta Education
- Memmi, Daniel; UQAM
- Neti, Saneeeta; University of Victoria
- Nickull, Duane; Adobe Systems, Inc.
- Ollivier-Gooch, Carl; The University of British Columbia
- Paulin, Michele; Concordia University
- Plaisant, Michel; University of Quebec
- Reid, Keith; Ontario Ministry of Agriculture
- Shewchenko, Nicholas; Biokinetics and Associates
- Steffan, Gregory; University of Toronto
- Vandenberghe, Christian; HEC Montreal

**Czech Republic**
- Kala, Zdenek; Brno University of Technology
- Korab, Vojtech; Brno University of technology
- Lhotska, Lenka; Czech Technical University

**Finland**
- Laodelma, Risto; University of Turku
- Salminen, Pekka; University of Jyvaskyla

**France**
- Cardey, Sylviane; University of Franche-Comté
- Klinger, Evelyne; LTCI – ENST, Paris
- Roche, Christophe; University of Savoie
- Valette, Robert; LAAS - CNRS

**Germany**
- Accorsi, Rafael; University of Freiburg
- Glatzer, Wolfgang; Goethe-University
- Gradmann, Stefan; Universitat Hamburg
- Groll, Andre; University of Siegen
- Klamma, Ralf; RWTH Aachen University
- Wurtz, Rolf P.; Ruhr-University Bochum

**India**
- Pareek, Deepak; Technology4Development
- Scaria, Vinod; Institute of Integrative Biology
- Shah, Mugdha; Mansukhlal Svayam

**Ireland**
- Eisenberg, Jacob; University College Dublin
- Feintuch, Uri; Hadassah-Hebrew University

**Italy**
- Badia, Leonardo; IMT Institute for Advanced Studies
- Berrettini, Maria; University of Palermo
- Carpaneto, Enrico; Politecnico di Torino

**Japan**
- Hattori, Yasunao; Shimane University
- Livingston, Paisley; Linghan University
- Srinivas, Hari; Global Development Research Center

**Netherlands**
- Mills, Melinda C.; University of Groningen
- Pires, Luiz Ferreira; University of Twente

**New Zealand**
- Anderson, Tim; Van Der Veen Institute

**Portugal**
- Cardoso, Jorge; University of Madeira
- Natividade, Eduard; Polytechnic Institute of Coimbra
- Oliveira, Eugenio; University of Porto

**Singapore**
- Tan, Fock-Lai; Nanyang Technological University

**South Korea**
- Kwon, Won Ho; Seoul National University

**Spain**
- Barrera, Juan Pablo Soto; University of Castilla
- Gonzalez, Evelio J.; University of La Laguna
- Perez, Juan Mendez; Universidad de La Laguna
- Royuela, Vicente; Universidad de Barcelona
- Vizcaino, Aurora; University of Castilla-La Mancha
- Villarasa, Celia Colombo; Open University of Catalonia

**Sweden**
- Johansson, Mats; Royal Institute of Technology

**Switzerland**
- Niinimaki, Marko; Helsinki Institute of Physics
- Pletka, Roman; AdNovum Informatik AG
- Rizzotti, Sven; University of Basel
- Specht, Matthias; University of Zurich

**Taiwan**
- Lin, Hsiung Cheng; Chienkuo Technology University
- Shyu, Yuh-Huei; Tamkang University
- Sue, Chauan-Ching; National Cheng Kung University

**United Kingdom**
- Ariwa, Ezendu; London Metropolitan University
- Biggam, John; Glasgow Caledonian University
- Coleman, Shirley; University of Newcastle
- Conole, Grainne; University of Southampton
- Dorfler, Viktor; Strathclyde University
- Engelmann, Dirk; University of London
- Eze, Emmanuel; University of Hull
- Forrester, John; Stockholm Environment Institute
- Jensen, Jens; STFC Rutherford Appleton Laboratory
- Kolovos, Dimitrios S.; The University of York
- McBurney, Peter; University of Liverpool
- Vetta, Adam; Oxford Brooks University
- Whyte, William Stewart; University of Leeds
- Xie, Changwen; Wicks and Wilson Limited

**USA**
- Bach, Eric; University of Wisconsin
- Bolzendahl, Catherine; University of California
- Bussler, Christoph; Cisco Systems, Inc.
- Charpentier, Michel; University of New Hampshire
- Chong, Stephen; Cornell University
- Collison, George; The Concord Consortium
- DeWeaver, Eric; University of Wisconsin - Madison
- Gans, Eric; University of California
- Gill, Sam; San Francisco State University
- Hunter, Lynette; University of California Davis
- Iceland, John; University of Maryland
- Kaplan, Samantha W.; University of Wisconsin
- Langou, Julien; The University of Tennessee
- Liu, Yuliang; Southern Illinois University Edwardsville
- Lok, Benjamin; University of Florida
- Minh, Chi Cao; Stanford University
- Morrissey, Robert; The University of Chicago
- Mui, Lik; Google, Inc
- Rizzotti, Albert; University of Southern California
- Rosenberg, Jonathan M.; University of Maryland
- Shaffer, Cliff; Virginia Tech
- Sherman, Elaine; Hofstra University
- Snyder, David F.; Texas State University
- Song, Zhe; University of Iowa
- Wei, Chen; Intelligent Automation, Inc.
- Yu, Zhiyi; University of California